Universidade do Minho
Escola de Engenharia

Duarte Filipe Oliveira Duque

Previsão e Identificação de Eventos de Quebra de Segurança em Vídeo-vigilância

Tese de Doutoramento em Tecnologias e Sistemas de Informação
Área de Sistemas de Computação e Comunicação

Trabalho efectuado sob a orientação de
Henrique Manuel Dinis dos Santos
Paulo Alexandre Ribeiro Cortez

Outubro de 2008

Declaração

Nome: Duarte Filipe Oliveira Duque

Endereço electrónico: duarteduque@dsi.uminho.pt T**elefone:** +351 965288059

Número do Bilhete de Identidade: 11356400

Título da Tese de Doutoramento:

Previsão e Identificação de Eventos de Quebra de Segurança em Vídeo-vigilância

Orientadores:

Henrique Manuel Dinis dos Santos

Paulo Alexandre Ribeiro Cortez

Ano de Conclusão: 2008

Ramo de Conhecimento do Doutoramento:

Tecnologias e Sistemas de Informação – Área de Sistemas de Computação e Comunicação

Universidade do Minho, 31 de Outubro de 2008

Assinatura: _____

Agradecimentos

Em primeiro lugar, gostaria de expressar um agradecimento muito especial ao Doutor Henrique Santos, por ter acompanhado e norteado a minha evolução a nível académico durante os últimos anos. Por ter sido um modelo a seguir, quer a nível científico, quer como pessoa.

Quero também reconhecer aqui o contributo fundamental do Doutor Paulo Cortez na elaboração deste doutoramento, cuja pronta resposta às minhas solicitações foram uma constante no decorrer deste trabalho.

Aos meus pais, Júlia e Hilário Duque, que sempre me apoiaram incondicionalmente nesta caminhada, muitas vezes com sacrifício, quero aqui deixar publicamente o meu reconhecimento.

Pretendo ainda agradecer aos meus amigos Miguel Ferreira e Pedro Gabriel pelo companheirismo, pelas intensas discussões e trocas de ideias que tive a oportunidade de partilhar, bem como a todos os meus amigos e colegas de laboratório, em especial à Fernanda Sarmento, ao Sérgio, Marco, Hélder, Filipe, Bete, Rosângela, Alysson e muitos outros que durante mais ou menos tempo partilharam comigo vivências que nos marcaram para sempre.

Agradeço o apoio concedido no âmbito do Programa de Bolsas de Formação Avançada da Fundação para a Ciência e a Tecnologia, através da concessão da bolsa de Doutoramento número SFRH / BD / 17259 / 2004, a qual financiou as propinas de doutoramento durante quatro anos.

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR Portugal

Aos meus Pais, Júlia e Hilário.

Resumo

Esta tese tem como propósito a detecção e previsão de comportamentos passíveis de originar uma quebra de segurança. Estes são reconhecidos por meio da observação de padrões de actividade humana, extraídos de sequências de imagens digitalizadas, adquiridas por intermédio de uma câmara de vídeo a cores, monocular e fixa. A aferição dos comportamentos é suportada pela informação obtida através da detecção, classificação e seguimento de objectos em movimento, minimizando a utilização de informação de contexto na cena observada e sem recurso a descrições de comportamentos previamente definidos.

De modo a atingir este objectivo, foram desenvolvidas técnicas de processamento e análise de imagem, associadas a métodos baseados em inteligência artificial para a modelação de padrões de comportamento. A segmentação de objectos em movimento foi assente numa abordagem de subtracção por plano de fundo adaptativo, com a capacidade de detecção de regiões da imagem afectadas por sombras e brilhos. Criou-se ainda um processo de remoção de *fantasmas*, i.e. falsas detecções observadas sempre que um objecto, pertencente ao plano de fundo, inicia um movimento de deslocação que o leva a abandonar o espaço anteriormente ocupado. O seguimento de objectos foi assegurado por uma técnica que recorre a *Modelos de Aparência*, e que possibilita o seguimento de objectos deformáveis, mostrando-se eficaz em situações de oclusão, fusão e separação de objectos. Para a detecção e previsão automática de comportamentos desenvolveram-se dois classificadores (*N-ary Trees* e *Dynamic Oriented Graph*) que, utilizando os dados provenientes das funções de processamento e análise de imagem, permitem modelar sequências temporais.

O sistema final, constituído pela junção dos múltiplos componentes propostos e implementado numa câmara de vídeo *inteligente*, foi testado com um conjunto de dados sintéticos, sendo posteriormente avaliado em ambiente real de vídeo-vigilância. Pela análise dos resultados experimentais, verificou-se que o sistema proposto permite realizar de forma eficaz a previsão de comportamentos de quebra de segurança.

Abstract

IDENTIFICATION AND FORECAST OF SECURITY BREAKS

IN VIDEO-SURVEILLANCE

This thesis has the purpose of detection and forecasting of behaviours susceptible to originate security breaks. These behaviours are recognized by means of human activity pattern observation, extracted from digital image sequences, acquired by a video colour camera, monocular and static. The assessment of the behaviours is supported by the information acquired through the detection, classification and tracking of moving objects, when minimizing the use of context information from the observed scene and without descriptions of previously defined behaviours.

In order to reach this goal, image processing and analysis techniques had been developed and associated with artificial intelligence methods for the behaviour pattern modelling. The segmentation of moving objects was based on an adaptive background subtraction approach capable of detecting regions of the image affected by shadows and highlights. A ghost's removal process was also developed, i.e. observed false detections whenever one object, pertaining to the background, initiates a movement that takes it to abandon the previously occupied space. The tracking of objects was assured by a technique that applies *Appearance Models*, which makes possible the tracking of deformable objects and reveals efficiency in situations of occlusion, fusion and splitting of objects. For the detection and automatic forecasting of behaviours, two classifiers (*N-ary Trees* and *Dynamic Oriented Graph*) were proposed. Both use preceding data from the processing and image analysis functions and they allow the modelling of temporal sequences.

The overall system, built from the junction of the components developed and implemented in an intelligent video camera, was tested with a synthetic dataset, being later evaluated in real environment of video-monitoring. The analysis of the experimental results has shown that the proposed system allows an efficient prediction of security break behaviours.

Índice

Siglas e Acrónimos

ART	*Adaptive Resonance Theory*
ATM	*Automated Teller Machine*
AUC	*Area Under the Curve*
BCLS	*Basic Competitive Learning Scheme*
BN	*Bayesian Network*
CAVIAR	*Context Aware Vision using Image-based Active Recognition*
CCD	*Charge-coupled Device*
CCTV	*Closed Circuit Television*
CIE	*Commission Internationale de L'éclairage*
CMOS	*Complementary Metal-oxide Semiconductor*
CNN	*Competitive Neural Network*
CNPD	Comissão Nacional de Protecção de Dados
CPU	*Central Processing Unit*
CRF	*Conditional Random Fields*
DARPA	*Defense Advanced Research Projects Agency*
DBN	*Dynamic Bayesian Network*
DFT	*Discrete Fourier Transform*
DOG	*Dynamic Oriented Graph*
DPE	*Distribuição da Potência Espectral*
DVR	*Digital Video Recorder*
DTW	*Dynamic Time Warping*
EM	*Expectation-Maximization*
FFT	*Fast Fourier Transform*
FN	Falso Negativo
FP	Falso Positivo
FSM	*Finite State Machine*
GM	*Gaussian Model*
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markov Model*
HSL	*Hue, Saturation and Luminance*
HSV	*Hue, Saturation and Value*
HVR	*Hybrid Video Recorder*
JPEG	*Joint Photographic Experts Group*
K-NN	*K-Nearest Neighbour*
LCSS	*Longest Common Subsequence*
LOTS	*Lehigh Omnidirectional Tracking System*
MEMM	*Maximum Entropy Markov Model*
MFA	*Marginal Fisher Analysis*
M-JPEG	*Motion JPEG*
MOG	*Mixture of Gaussians*
MPEG	*Moving Picture Experts Group*
MRF	*Markov Random Field*
NIHC	*Numeric Iterative Hierarchical Cluster*
NN	*Neural Network*
NTSC	*National Television System Committee*
NVR	*Network Video Recorder*
PAL	*Phase Alternating Line*

PCA	*Principal Components Analysis*
PETS	*Performance Evaluation of Tracking and Surveillance*
PTZ	*Pan, Tilt and Zoom*
RGB	*Red, Green and Blue*
SAFEE	*Security of Aircraft in the Future European Environment*
SOM	*Self-Organizing Map*
SPU	*Sensor Processing Unit*
STUM	Serviços Técnicos da Universidade do Minho
SVM	*Support Vector Machine*
VIEWS	*Visual Inspection and Evaluation of Wide-area Scenes*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VQ	*Vector Quantization*
VSAM	*Video Surveillance and Monitoring*
W4	*Who? When? Where? What?*
XML	*Extensible Markup Language*

Índice de Figuras

Índice de Tabelas

Capítulo 1

1. Introdução

O presente capítulo tem como propósito apresentar o contexto em que este trabalho de doutoramento se insere, bem como os motivos que levaram à elaboração desta tese. Ainda durante o capítulo introdutório serão identificados os principais contributos e objectivos, assim como a metodologia adoptada para a elaboração deste trabalho.

A Secção 1.1 descreve, de uma forma sucinta, a evolução dos sistemas de vídeo-vigilância, desde o seu surgimento até à actualidade. Pretende-se, deste modo, proporcionar ao leitor uma breve resenha histórica sobre o desenvolvimento tecnológico ocorrido nos sistemas de vídeo-vigilância, área em que este trabalho se enquadra. Na Secção 1.2 apresenta-se a motivação que levou à elaboração desta tese. Assim, descrevem-se alguns estudos que identificam as fragilidades e limitações dos actuais sistemas e aponta-se uma proposta de solução para os problemas levantados. Os principais objectivos e contributos são identificados na Secção 1.3, enquanto que na Secção 1.4 se descreve a metodologia adoptada para a elaboração deste trabalho. Por fim, na Secção 1.5 é sumariamente apresentada a organização geral da dissertação.

1.1. Enquadramento

A manutenção da segurança de espaços públicos e privados tem como finalidade a detecção e prevenção de eventos não autorizados. Apesar de na maioria dos casos essa segurança ser mantida com o recurso a agentes de segurança dispostos no terreno (polícia ou funcionários de segurança privada), em determinadas situações a relação custo/benefício atinge valores proibitivos. Nestes casos, o recurso a equipamentos de vídeo-vigilância permite uma melhor gestão dos recursos humanos, através de uma vigilância remota.

A utilização de sistemas de vídeo-vigilância disponibiliza ainda benefícios adicionais, uma vez que estes permitem a investigação e análise dos eventos após a sua ocorrência. Como exemplo, pode-se verificar o impacto que os sistemas de vídeo-vigilância tiveram nas investigações dos atentados bombistas de Londres em Julho de 2005, ou no 11 de Setembro de 2001 nos Estados Unidos da América.

Em consequência das vantagens inerentes à utilização de sistemas de vídeo-vigilância na manutenção da segurança, estes têm sido alvos de intensa investigação, em especial nas duas últimas décadas. Como resultado, verificou-se uma crescente evolução, quer a nível dos subsistemas que o compõem, quer da sua arquitectura. Do ponto de vista histórico, pode-se afirmar que essa evolução resultou em três gerações de sistemas.

Na primeira geração de sistemas de vídeo-vigilância (1960-1980) [Hyder et al., 2002], a aquisição de imagens, a sua transmissão, processamento e armazenamento suportavam-se em tecnologias puramente analógicas. O objectivo consistia em centralizar numa sala de controlo as imagens captadas por diversas câmaras de vídeo, que se encontravam distribuídas num edifício ou espaço público. Este tipo de sistema pode ser visto como uma extensão do olho humano, no ponto de vista espacial, em que um operador analisa sequências de vídeo através de um conjunto de monitores, onde as cenas monitorizadas pelas múltiplas câmaras são multiplexadas e apresentadas periodicamente e por ordem predefinida.

Apesar de útil, esta abordagem à vídeo-vigilância apresentava várias desvantagens. Devido à natureza analógica do sinal de vídeo disponibilizado pelas câmaras, a sua transmissão acarretava diversos problemas. Entre os factores negativos desta tecnologia encontravam-se: o custo elevado da cablagem; as curtas distâncias suportadas entre as câmaras e o

equipamento de monitorização e armazenamento; e a degradação da qualidade do vídeo devido ao ruído induzido na cablagem [Sacchi et al., 1999]. O arquivamento e consulta de eventos constituíam também um factor negativo, uma vez que estes eram armazenados num grande número de cassetes de vídeo, e a sua detecção estava dependente apenas da atenção de um operador humano [Donald, 1999].

A evolução seguiu o rumo dos sistemas híbridos. Assim, a segunda geração de sistemas de vídeo-vigilância, que surgiu entre a década de 80 e perdurou até meados do ano 2000, integrava subsistemas analógicos e digitais de modo a suprimir algumas das desvantagens dos seus predecessores. A transição de analógico para digital verificou-se de forma ténue na transmissão do sinal, com o aparecimento de conversores de sinal de vídeo de analógico para digital e de digital para analógico. Utilizando estas técnicas, era possível aumentar de forma significativa a distância entre as câmaras e o equipamento receptor, bem como eliminar a indução de ruído no sinal de vídeo e a sua consequente degradação. Contudo, a evolução mais significativa ocorreu nos sistemas de aquisição, tratamento e armazenamento de imagem, que passaram a ser baseados em computador. Surgem deste modo os *DVRs*[1], que fazendo uso das suas capacidades de processamento implementam o armazenamento de vídeo comprimido (*JPEG*, *M-JPEG*, *MPEG*, entre outros) em discos rígidos, suprimindo assim o problema da degradação da qualidade do vídeo que ocorria com a utilização das fitas magnéticas.

A utilização de sistemas baseados em computador trouxe outros benefícios como a gestão automática do vídeo armazenado no sistema e a transmissão de vídeo digital em tempo-real sobre redes *ethernet*. É ainda nesta altura que são dados os primeiros passos na detecção automática de movimento, surgindo diversos produtos que implementam esta nova funcionalidade.

Em paralelo com a crescente popularidade e aceitação, por parte do mercado, deste novo tipo de soluções de vigilância, também na comunidade científica se verificou um notório aumento do interesse e esforço de investigação em áreas como o processamento e análise de imagem. Como resultado, enormes progressos foram alcançados, acrescentando novas funcionalidades aos sistemas existentes. Assim, durante esse período de forte expansão tecnológica, assistiu-se ao surgimento de novas técnicas de compressão de vídeo, à

[1] Do inglês, *Digital Video Recorder*.

transmissão de áudio e vídeo em tempo-real, e à segmentação, identificação e seguimento de objectos em cenas complexas.

A terceira geração de sistemas de vídeo-vigilância surge a partir do ano 2000 [Hyder et al., 2002], completando a migração para digital dos sistemas precedentes. Nos equipamentos desta geração, o sinal de vídeo é convertido para o domínio digital nas próprias câmaras (digitais), que possuem já consideráveis capacidades de processamento. O vídeo é então submetido a operações de processamento de imagem de baixo nível e comprimido de forma a ser transmitido via *ethernet* para um ou vários computadores clientes. Deste modo, a largura de banda requerida para a transmissão do vídeo é minimizada, de acordo com as especificações de qualidade de imagem definida pelo utilizador do sistema. Outras funcionalidades anteriormente implementadas pelos *DVRs* migraram também para as câmaras, como é o caso da detecção automática de movimento. Este tipo de abordagem tem como principal vantagem permitir a utilização das infra-estruturas de rede existentes, reduzindo desta forma os custos de instalação deste tipo de soluções.

A tendência de evolução dos sistemas de vídeo-vigilância aponta para o paradigma da migração de um sistema centralizado para um sistema de vigilância distribuído. A principal motivação para esta migração tem por base a crescente funcionalidade, disponibilidade e autonomia deste tipo de sistema. Como exemplo, através de um sistema de vídeo-vigilância distribuído, um operador poderia monitorizar simultaneamente um conjunto de espaços, fisicamente distantes, utilizando as infra-estruturas de rede informática partilhadas por outros dispositivos.

Atendendo aos factos apresentados, parece claro que qualquer inovação que venha a complementar as soluções existentes deve ter em consideração as tendências da indústria da vídeo-vigilância, tendo em atenção os recursos computacionais existentes bem como os requisitos em termos temporais. A nova geração de sistemas de vídeo-vigilância aponta então para a utilização de câmaras inteligentes, com capacidade não só de processar, mas também de analisar sequências de imagens. A utilização, por parte destas novas câmaras, de sistemas operativos livres e abertos como o *GNU/Linux*, e a possibilidade, que alguns dos fabricantes já disponibilizam aos utilizadores, de adicionar novas funcionalidades ao sistema, tornam este tipo de tecnologia o campo ideal para teste de novas soluções.

1.2. Motivação

A vídeo-vigilância tem como propósito evitar que determinadas situações degenerem em crimes, através da detecção de actividades suspeitas, causando alarmes ou acções similares com fins preventivos. Tipicamente, a vigilância de um espaço com recurso a imagens de vídeo é assegurada através de um circuito fechado de televisão (*CCTV*[2]), monitorizado por um profissional de segurança. Com frequência, um único operador de vídeo-vigilância tem a seu cargo a monitorização de um elevado número de câmaras.

No início da década de 70, investigadores da Unidade de Psicologia Aplicada da Universidade de Cambridge analisaram o desempenho do elemento humano no sistema de vídeo-vigilância. Pretendia-se identificar a eficácia do operador na detecção de eventos bem como os factores que poderiam influenciar o seu desempenho.

Os resultados destes estudos [Tickner & Poulton, 1972; 1973] demonstraram que, monitorizando até nove câmaras em simultâneo, em média, apenas 83% dos eventos anormais eram detectados. Contudo, para uma monitorização de dezasseis câmaras de vídeo, a média de eventos detectados decai de forma significativa para 64%. Os mesmos estudos apontaram ainda para: a degradação dos níveis de atenção e eficácia na detecção de eventos anormais ao longo do tempo; a relevância de factores como a idade e o sexo do supervisor humano no seu desempenho; e para a importância da resolução das imagens fornecidas ao operador para uma maximização do número de incidentes detectados.

Recentes testes militares [Ainsworth, 2004] sobre a resposta dos operadores vieram realçar as limitações dos actuais sistemas de vídeo-vigilância e reforçar a necessidade do desenvolvimento de mecanismos de auxílio nas tarefas de vigilância. Tais conclusões devem-se principalmente à verificação de que, visualizando apenas dois monitores, um operador perde 45% da acção após doze minutos de monitorização contínua, e após vinte e dois minutos perde 95% da acção.

Nos actuais sistemas de vídeo-vigilância, apesar de se almejar uma monitorização de comportamentos com vista à actuação preventiva em caso de quebra das normas definidas, a segurança é sobretudo passiva. O operador apenas é alertado após a ocorrência de um

[2] Do inglês, *Closed Circuit Television*.

evento de quebra de segurança, através de sensores de intrusão ou, em sistemas de vídeo-vigilância digitais, através de detecção de movimento em zonas restritas. A informação armazenada neste tipo de sistemas só será requisitada para análise, após essa detecção.

Apesar de os sistemas de vídeo-vigilância digitais tirarem partido dos seus recursos computacionais em tarefas de armazenamento de imagem, efectuando a gravação por detecção de movimento, compressão de vídeo, e oferecerem aos utilizadores poderosos processos de consulta de gravações de vídeo, estes continuam a facultar essencialmente a mesma funcionalidade dos antecessores. Ao operador do sistema exige-se a atenção permanente para que, em função do nível de segurança definido, reaja a situações de alarme visualmente detectadas.

Existe um potencial de evolução, no sentido de auxiliar os operadores humanos em tarefas de vigilância, proporcionando assim uma segurança activa. O objectivo será então o de detectar e prever automaticamente comportamentos passíveis de originar eventos de quebra de segurança.

Os estudos apresentados nos parágrafos anteriores desta secção apontaram para a debilidade da tarefa de monitorização e detecção de eventos em tempo-real por operadores humanos, que à partida se poderia pensar ser trivial e de elevada taxa de sucesso. Por conseguinte, serão os operadores capazes de realizar com êxito uma tarefa mais complexa, i.e., a previsão de comportamentos observando sequências de imagens obtidas por um sistema de *CCTV*?

Um estudo [Troscianko et al., 2004] realizado pelo Departamento de Psicologia Experimental da Universidade de Bristol, em conjunto com a Escola de Psicologia da Universidade de Massey e com o Departamento de Psicologia da Universidade de Sussex, concluiu que é possível aos humanos prever comportamentos perigosos através de sequências de imagens de *CCTV*. Neste estudo, em que imagens previamente capturadas eram fornecidas a um observador, utilizando um único monitor, verificou-se que essa tarefa poderia ser realizada com uma taxa de sucesso de cerca de 81%. O mesmo estudo apontava um outro dado importante, isto é, que não existia uma diferença significativa entre peritos de vigilância e pessoas comuns. Ambos os grupos obtinham taxas de sucesso semelhantes.

Constata-se assim que o desenvolvimento de mecanismos que automatizem o processo de análise das imagens vídeo, auxiliando o vigilante na tarefa de detecção de comportamentos

anormais, é uma necessidade real e vital para o sector de serviços de vídeo-vigilância. Como resposta a esta necessidade, no âmbito desta tese propunha-se o desenvolvimento de um sistema baseado em computador que, através de técnicas de processamento e análise de sequências de imagens em formato digital, e utilizando métodos de classificação baseados em técnicas de *Data Mining*, fosse capaz de detectar e interpretar movimento, de forma a identificar e prever eventos de quebra de segurança.

1.3. Objectivos e Contributos

O objectivo principal definido para este trabalho de doutoramento consistiu na investigação de técnicas e metodologias que permitissem detectar e prever comportamentos anómalos, passíveis de originar eventos de quebra de segurança, em locais sob vigilância. Essa análise de comportamentos deriva da observação de padrões de actividade humana, extraídos de sequências de imagens digitalizadas, recorrendo a técnicas de *Data Mining*. A aferição do tipo de comportamento foi executada de modo a minimizar a necessidade de informação de contexto da cena observada ou de comportamentos previamente definidos. As imagens utilizadas pelo sistema são obtidas através de uma câmara de vídeo a cores, monocular e fixa, com vista à integração dos resultados desta investigação em sistemas de vídeo-vigilância comerciais.

Assim, esta tese contribui com:

(i) uma nova técnica de segmentação de objectos em movimento, baseada na combinação de subtracção de plano de fundo adaptativo com a diferença entre imagens (a técnica proposta permite segmentar objectos, removendo do resultado da segmentação erros de sombras, brilhos e fantasmas) [Duque et al., 2005; 2008];

(ii) um conjunto de técnicas de extracção de dados, nomeadamente para classificação dos objectos observados em pessoas, grupos de pessoas e veículos [Duque et al., 2006b];

(iii) um algoritmo de seguimento de objectos deformáveis, que garante o seguimento durante a ocorrência de interacções complexas entre objectos, como são o caso da fusão, separação e oclusão de objectos [Duque et al., 2006b; 2008]; e

(iv) duas propostas originais para a identificação e previsão de padrões de actividade, baseadas em dois classificadores denominados por *N-ary Trees* e *Dynamic Oriented Graph* [Duque et al., 2006; 2007; 2007b].

1.4. Metodologia

Com esta tese pretende-se comprovar a hipótese da previsão automática de comportamentos anormais em espaços públicos, pelo recurso a imagens capturadas por uma câmara de vídeo. Trata-se de um problema complexo cuja validação de uma eventual solução implica uma verificação metódica dos diversos componentes que a constituem. Como tal, propõe-se uma abordagem que decompõe o problema nos seguintes elementos de menor complexidade: segmentação de objectos em movimento; seguimento de objectos; e previsão de comportamentos.

A segmentação e o seguimento de objectos em movimento são dois temas bastante estudados pela comunidade científica, constituindo a base de todo o sistema. É pela aplicação de técnicas de processamento e análise de imagem que se torna possível a extracção das sequências de dados que caracterizam as actividades de cada objecto em cena. Em resultado da necessidade de uma plataforma comum de suporte à avaliação e análise comparativa dos resultados da investigação nestas áreas, a *Performance Evaluation of Tracking and Surveillance Workshop* [PETS, 2001] propôs um conjunto de dados de treino e de teste. Este conjunto de dados foi empregue na avaliação e validação das técnicas de segmentação e seguimento de objectos, desenvolvidas no âmbito deste trabalho.

As sequências de dados são utilizadas na modelação de padrões de comportamento com vista à identificação e previsão de novas observações. A avaliação das técnicas desenvolvidas para o efeito foi realizada com recurso a um conjunto de dados sintéticos, que simulam trajectórias de objectos num determinado espaço. A técnica que demonstrou melhor desempenho foi então seleccionada para inclusão no protótipo do sistema de detecção de comportamentos.

O teste final do sistema, que consistiu uma prova de conceito, foi levado a cabo no campus da Universidade do Minho, em Guimarães. Com essa finalidade, definiu-se uma estratégia de avaliação do sistema em ambiente real, que implicou a monitorização de uma área do campus durante um alargado período de tempo de modo a garantir uma aprendizagem dos padrões de comportamento. Os modelos de comportamento assimilados durante esse período foram utilizados para testar a capacidade de identificação e previsão do sistema. Para este efeito utilizaram-se "actores" cuja função consistia na violação de uma área restrita pertencente ao espaço monitorizado.

A execução do trabalho apresentado consistiu em cinco fases distintas, que aqui se descrevem sucintamente:

Fase 1 – Revisão de literatura existente sobre segmentação e seguimento de objectos em movimento. Levantamento e estudo de bibliografia subordinada à temática da detecção e previsão automática de comportamentos, utilizando dados recolhidos por câmaras de vídeo.

Fase 2 – Análise comparativa dos métodos de segmentação existentes. Levantamento das suas vantagens e desvantagens. Definição de requisitos para a segmentação de objectos. Estudo de técnicas de detecção e remoção de sombras e brilhos. Desenvolvimento e avaliação de um algoritmo de segmentação.

Fase 3 – Estudo comparativo dos métodos de seguimento de objectos identificados no levantamento bibliográfico. Descrição dos requisitos para o seguimento de objectos. Implementação e avaliação de um algoritmo.

Fase 4 – Análise das abordagens existentes para identificação e previsão de comportamentos. Caracterização de eventos anormais do ponto de vista de um sistema de vídeo-vigilância e identificação dos atributos relevantes passíveis de serem extraídos do processo de segmentação e seguimento de objectos. Desenvolvimento de *software* de apoio à produção de dados de seguimento sintéticos. Proposta e desenvolvimento de uma nova abordagem ao problema. Análise de desempenho da abordagem proposta com dados sintéticos.

Fase 5 – Construção de um protótipo, constituído por uma câmara de vídeo digital a cores, com captura por esquadrinhamento progressivo. Teste do protótipo em situação real de forma a realizar a prova de conceito.

1.5. Organização da Dissertação

Esta tese encontra-se estruturada em sete capítulos. No capítulo inicial, apresenta-se uma introdução geral ao problema em consideração. São apresentados os contributos e objectivos, o enquadramento da tese e a metodologia adoptada. O Capítulo 2 é iniciado por uma revisão do trabalho relacionado com a segmentação e seguimento de objectos em movimento. Prossegue-se com um levantamento do "estado da arte" na detecção e previsão automática de comportamentos observados por câmaras de vídeo.

O terceiro capítulo foca-se na problemática da representação de cor em imagem digital. Neste contexto, é apresentado o modelo de reflexão dicromático, assim como um conjunto de espaços de cor invariantes com relevo para a problemática da segmentação.

O Capítulo 4 aborda a segmentação de objectos em movimento. Neste âmbito, descreve-se uma técnica para a detecção e remoção de sombras e brilhos, associada a um mecanismo para detecção de *fantasmas*. O capítulo encerra com a avaliação da técnica de segmentação proposta.

O seguimento de objectos é tratado no quinto capítulo, onde se apresenta uma técnica baseada em *Modelos de Aparência*. A técnica proposta é avaliada com recurso a um conjunto de dados de teste fornecidos pela PETS [PETS, 2001b]. Neste capítulo é ainda referido o método desenvolvido para a classificação dos objectos em uma de três classes: pessoas, grupos de pessoas, e veículos.

O sexto capítulo é dedicado à detecção e previsão de comportamentos, onde se descreve a abordagem pretendida para a classificação de comportamentos. São apresentados dois novos classificadores: o *N-ary Trees* e o *Dynamic Oriented Graph*. Estes classificadores são testados numa primeira fase sobre dados sintéticos, de modo a examinar o seu desempenho. O classificador mais apto (*Dynamic Oriented Graph*) foi então aplicado em condições reais.

A tese termina com a apresentação das conclusões (Capítulo 7), onde também se referem algumas linhas de orientação para trabalho futuro.

Capítulo 2

2. Trabalho Relacionado

Pretende-se neste capítulo apresentar uma visão geral do "estado da arte" das áreas temáticas subjacentes a esta tese. Assim, em primeiro lugar, apresentam-se os principais projectos e contributos, por ordem cronológica, para a segmentação e seguimento de objectos em movimento, observados por câmaras de vídeo monoculares e fixas. Este levantamento é complementado por uma análise comparativa das propostas apresentadas.

Seguidamente, enunciam-se os esforços mais relevantes, empreendidos pela comunidade científica, no sentido de detectar e prever comportamentos, tendo por base dados recolhidos por meio de técnicas de processamento e análise de imagem. Neste levantamento, pretende-se dar uma visão do percurso evolutivo das técnicas de classificação e previsão de comportamentos, desde os finais da década de oitenta até à actualidade. Finaliza-se com uma comparação das diversas abordagens apresentadas, indicando as vantagens e inconvenientes de cada alternativa.

2.1. Segmentação e Seguimento de Objectos em Movimento

O desafio de detectar e seguir objectos em movimento tem sido abordado por vários investigadores que, naturalmente, tentam alcançar o melhor desempenho na realização desta tarefa. A detecção de objectos em movimento é um problema complexo quando lida com ambientes não controlados, como é o caso de cenas exteriores. Em tais cenários, sombras, alterações da intensidade da luz, deformação e interacção entre objectos, são problemas bastante comuns, que habitualmente constituem uma fonte de ruído no sistema.

Várias abordagens foram propostas. Estas divergem em diversos aspectos, principiando pelo tipo de aquisição de imagem que é efectuado. Neste domínio, podem-se encontrar soluções que recorrem à utilização de uma única câmara, fixa ou móvel, enquanto outras aplicam visão estéreo. Em alguns sistemas as imagens são adquiridas em formato de escala de cinzentos, em detrimento da componente de cor que é explorada por outras abordagens.

O leque de soluções propostas para a segmentação de objectos em movimento é também muito vasto. As principais variantes incluem a diferenciação temporal [Otsu, 1979; Pun, 1980], o fluxo óptico [Horn & Schunck, 1981] e a subtracção do plano de fundo [Haritaoglu et al., 2000]. Esta última abordagem é a mais recorrente, compreendendo um conjunto de variantes, e.g. gaussiano, mistura de gaussianos e distribuição bimodal.

Os sistemas que têm vindo a ser anunciados na comunidade científica não apresentam uma uniformidade no que respeita à abordagem a adoptar para o seguimento de objectos. Embora seja uma área mais recente que a segmentação, existe já um grande número de técnicas ensaiadas para a resolução deste problema. Métodos que utilizam *Hidden Markov Models* [Rabiner, 1989], Redes *Bayesianas* [Duda et al., 2001], e Filtros *Kalman* [Stauffer & Grimson, 2000] são exemplos de soluções propostas.

Em poucos casos, os problemas da segmentação e seguimento de objectos são considerados como um só, como proposto no trabalho de Yamamoto [Yamamoto et al., 1995], onde o seguimento de múltiplos objectos em movimento é efectuado com base num fluxo óptico, combinado com informação espacial. Neste sistema, as imagens adquiridas em escala de cinzentos são processadas de forma a calcular os vectores do fluxo óptico. O cálculo do fluxo óptico por ser uma tarefa que exige enormes recursos computacionais é assegurado por um processador de imagem especialmente concebido para o efeito [Shioara et al., 1993].

Os autores deste trabalho assumem que os pontos correspondentes a um determinado objecto possuem vectores de fluxo similares. Assim, a segmentação dos objectos é efectuada através da extracção de regiões conectadas que encerrem vectores de fluxo óptico idênticos. Este tipo de solução, para além da desvantagem de necessitar de elevados recursos computacionais, nomeadamente do hardware dedicado, possui a forte limitação do pressuposto de que num objecto em movimento todas as partes que o constituem se movimentam num mesmo sentido. Este pressuposto só é válido no caso de corpos rígidos, e apenas quando não se verifica um movimento de rotação. Por esta razão, a técnica proposta por Yamamoto não é a indicada para a detecção e o seguimento de pessoas.

Para além disso, o fluxo óptico pressupõe uma iluminação constante ao longo do tempo. Esta condição não se verifica em situações reais, onde frequentemente ocorrem variações da intensidade luminosa decorrentes de fenómenos atmosféricos (e.g. nuvens que atenuam a incidência dos raios solares sobre os objectos em cena). Assim, técnicas de segmentação de objectos baseadas em fluxo óptico carecem de universalidade de utilização, sendo aplicáveis apenas em ambientes controlados.

Em 1997, a *Defense Advanced Research Projects Agency* (*DARPA*) iniciou um projecto de três anos, denominado *Video Surveillance and Monitoring* (*VSAM*) [Stauffer & Grimson, 1999; 2000; Collins et al., 2000]. No âmbito desse programa, foi desenvolvido um sistema com o objectivo de recolher e disseminar, de forma automática, informação em tempo-real obtida a partir de um campo de batalha. O sistema era composto por um conjunto de sensores que incluíam câmaras a cores fixas, câmaras *PTZ*[3] e sensores térmicos.

A arquitectura deste sistema definia a conexão de cada câmara a uma unidade de processamento, denominada *Sensor Processing Unit* (*SPU*), que teria a seu cargo o processamento do sinal de vídeo. Em cada uma destas unidades *SPU*, a segmentação de objectos era efectuada com base em dois processos: análise de ponto e análise de região. A análise de ponto tinha como objectivo determinar o seu estado. Este poderia ser estacionário ou transitório. Tal análise era efectuada através da observação da intensidade luminosa do ponto ao longo do tempo. Por outro lado, a análise de região trabalhava com o agrupamento de pontos em regiões de movimento ou em regiões estáticas.

[3] Câmara de vídeo com capacidade de rotação horizontal e vertical e que permite a variação da distância focal.

O sistema *VSAM* foi desenvolvido de modo a suportar a detecção de objectos durante a ocorrência de oclusões parciais. Este propósito foi alcançado graças a um esquema de segmentação por camadas. O seguimento de objectos era assegurado por uma variante dos Filtros *Kalman* que incorporava a correlação de imagens.

A classificação dos objectos segmentados foi efectuada com recurso a redes neuronais. Para o efeito utilizou-se uma rede com três camadas (quatro entradas, dezasseis neurónios na camada escondida e três saídas), treinada através de um algoritmo de *backpropagation*. Dos quatro parâmetros de entrada faziam parte: a área do objecto, em pontos; a razão do quadrado do perímetro pela área; a proporção entre a altura e a largura da caixa delimitadora do objecto; e o *zoom* da câmara. A rede classificava então cada objecto como: humano; múltiplos humanos; ou veículo.

O *VSAM* apresentava, contudo, algumas limitações. Nomeadamente, no que diz respeito à técnica de segmentação utilizada pelo sistema, uma vez que esta não se encontrava dotada de mecanismos de detecção e remoção de sombras. Deste modo, o sucesso do processo de segmentação estaria dependente da não ocorrência destes fenómenos.

Baseado num mecanismo de subtracção de plano de fundo adaptativo, o algoritmo de segmentação usado no *VSAM* era susceptível de ser influenciado por fantasmas, i.e. falsos positivos originados pela deslocação de objectos pertencentes ao plano de fundo. Este facto implicava que durante a inicialização do sistema não existissem objectos em cena.

O W^4 (*Who? When? Where? What?*) [Haritaoglu et al., 1998; 2000], foi um sistema de vídeo-vigilância com características de funcionamento em tempo real, desenvolvido para detectar e seguir simultaneamente múltiplas pessoas em ambientes exteriores. Este sistema operava sobre imagens monoculares em escala de cinzentos, ou em imagens obtidas através de câmaras de infra-vermelhos. A segmentação de objectos em movimento foi realizada por um algoritmo de subtracção do plano de fundo adaptativo, onde a imagem de referência é modelada por uma distribuição bimodal, construída por outros valores estatísticos do plano de fundo, gerados durante um período de treino.

Neste sistema, os objectos eram seguidos com recurso a um molde de textura temporal aplicado em conjunto com a informação da forma do objecto. O W^4 possuía a capacidade de distinguir pessoas dos restantes objectos e de efectuar o seguimento de cada indivíduo após a ocorrência de uma oclusão. Este sistema tinha ainda a aptidão para detectar, no caso

de se tratar de uma pessoa isolada, seis partes do corpo (cabeça, mãos, pés e tronco) e determinar se essa pessoa transportava ou não um objecto.

Apesar das vantagens apontadas a este sistema, o *W*^{4} não efectuava a detecção de sombras. Este facto originava que, em situações reais, onde as sombras são um factor indutor de deformações no resultado do processo de segmentação, todas as técnicas aplicadas a jusante produzam resultados errados. A não utilização da informação da cor na definição da textura dos objectos, tornava o processo de seguimento de objectos pouco robusto. Por exemplo, se dois objectos com a mesma forma se intersectavam, sendo um deles vermelho e o outro azul, o sistema poderia não efectuar correctamente o seguimento destes objectos, uma vez que, no espaço de cor de cinzentos os objectos teriam uma tonalidade semelhante.

O sistema *Sakbot* [Cucchiara et al., 2001] também faz uso de um algoritmo adaptativo de subtracção do plano de fundo, para segmentar regiões de primeiro plano referentes a objectos em movimento. Para este sistema, foi desenvolvido um método de actualização do plano de fundo, denominado *Sε&KB*. Este era utilizado em conjunto com um algoritmo de detecção de sombras de forma a se obter um resultado de segmentação mais preciso.

O *Sakbot* fazia já uso da informação de cor das imagens adquiridas. O sistema efectua a segmentação dos objectos através do cálculo do valor da diferença absoluta entre a imagem actual e a imagem de referência, para cada ponto e para cada componente de cor no espaço *RGB*. O valor máximo obtido para cada ponto é posteriormente confrontado com dois valores de comparação, um valor mínimo e um valor máximo, efectuando-se desta forma uma segmentação com histerese.

As regiões de primeiro plano detectadas pelo processo anterior, passavam então por um método de filtragem, que removia as regiões com área considerada desprezável. Nas regiões resultantes era calculado o fluxo óptico médio. Nos casos em que este fosse menor que um predeterminado valor, essa região era considerada como um fantasma e removida do resultado da segmentação. Contudo, como já foi referido anteriormente, as técnicas de fluxo óptico evidenciam algumas limitações, nomeadamente no que diz respeito ao pressuposto da existência de uma luminosidade constante.

O sistema *Sakbot* possuía mecanismos para detecção de sombras que trabalhavam no espaço de cor *HSV*. Este facto permitia uma segmentação dos objectos mais precisa do que a obtida com os sistemas anteriores, e aplicável em situações reais.

A abordagem adoptada no *Sakbot* para o seguimento de objectos consistia num sistema de produção de regras em cadeia, formalizando o ambiente conhecido e as relações entre os objectos extraídos. Este era um sistema bastante simples que não permitia detectar individualmente objectos num grupo, nem durante a ocorrência de oclusões. O *Sakbot*, apesar de possuir mecanismos que lhe permitiam obter resultados de segmentação com elevada precisão, padecia de um fraco desempenho no processo de seguimento de objectos, que evidenciou numerosas limitações.

Dan Buzan [2004] descreve um sistema que efectua a detecção e seguimento de objectos e realiza o agrupamento de trajectórias similares. Este sistema modelava o plano de fundo como uma distribuição gaussiana. Para tal, aquando do arranque do sistema, era efectuada a aprendizagem da imagem de referência, durante um período de tempo que variava de acordo com o tipo de ambiente observado. Posteriormente, o plano de fundo era então actualizado através de um filtro adaptativo.

O sistema proposto por Buzan efectuava o seguimento de objectos em movimento, baseado numa técnica suportada em Filtros *Kalman* Estendidos. Todavia, o autor apontou algumas falhas no seu sistema. Uma das limitações manifestava-se quando, numa oclusão entre dois ou mais objectos, os perseguidores não se encontravam suficientemente estabilizados, podendo desta forma perder o rasto dos objectos. Um outro caso em que o sistema experimenta dificuldades ocorre sempre que um objecto desaparece temporariamente da cena, quer seja devido à oclusão por elementos do plano de fundo, ou pela similaridade de cor com o plano de fundo. Neste caso o sistema indica o objecto como desaparecido, após este não ter sido detectado em mais de três imagens consecutivas.

No projecto *CAVIAR* [Hall et al., 2005], a segmentação de regiões de primeiro plano é efectuada pelo método apresentado no *Lehigh Omnidirectional Tracking System* (*LOTS*), proposto em [Boult et al., 1998; 1999; 2001]. Este método inclui a modelação adaptativa de plano de fundo múltiplo, uma nova abordagem para o agrupamento espaço-temporal denominada *Quasi-connected Components* [Boult et al., 2001], e a subtracção do plano de fundo por *threshold* adaptativo.

Neste projecto, os objectos são seguidos por meio de Redes *Bayesianas* como descrito em [Jorge et al., 2004]. Esta nova variante permite ao sistema trabalhar sobre sequências de imagens de longa duração, o que não acontecia com os anteriores modelos baseados neste tipo de redes [Abrantes et al., 2002]. Tal é conseguido através de um processo que reduz

gradualmente a influência da informação passada na tomada de decisões, evitando assim a explosão de combinações e mantendo a complexidade da rede em limites razoáveis. Contudo, os resultados experimentais apresentados até à data são bastante escassos (em [Jorge et al., 2004] o conjunto de dados possuía apenas trinta e quatro trajectórias), não permitindo obter uma análise fiável do desempenho do sistema.

2.1.1. Avaliação da Segmentação e Seguimento de Objectos

Como se pode verificar, os sistemas propostos até à data divergem de forma significativa no que respeita às escolhas das abordagens de segmentação e de seguimento de objectos. A questão que surge é saber como testar as soluções apresentadas, de forma a avaliar o seu desempenho de um modo quantitativo. Para esse efeito é necessário testar os diferentes sistemas com um mesmo conjunto de dados, i.e. uma sequência de imagens comum, e posteriormente efectuar a medição do desempenho com um conjunto de métricas que permitam comparar as diferentes técnicas.

O problema da avaliação de desempenho dos sistemas de seguimento não é novo, tendo sido já abordado por alguns autores. A motivação destes investigadores era exactamente a necessidade de medir o progresso da sua investigação e comparar as diferentes abordagens ao problema. Assim, no ano de 2001, a *Performance Evaluation of Tracking and Surveillance Workshop* (*PETS*) [PETS, 2001] propôs um conjunto de sequências de imagens digitalizadas e definiu um formato de saída de dados em XML para os resultados do processo de seguimento.

Para avaliar o desempenho de um sistema de segmentação e de seguimento é necessário, em primeiro lugar, gerar os dados de referência a partir das sequências de imagens de teste. Este é um processo manual, onde é necessário que um operador humano defina as caixas delimitadoras de cada objecto de interesse, ou seleccione os pontos da imagem que pertencem a cada objecto em movimento. A geração manual destes dados de referência revela-se extremamente custosa em termos temporais, apesar da existência de algumas ferramentas de suporte, e.g. *ViPER* [Doermann & Mihalcik, 2000] e *ODViS* [Jaynes et al., 2002].

Pelo facto da geração dos dados de referência ser efectuada por um operador humano, estes poderão não ser estritamente correctos. Na verdade é provável que sejam afectados por erros sistemáticos. Isto porque a noção de limite dos objectos pode variar de operador

para operador provocando assim, de uma forma sistemática, erros nos dados de referência. Deste modo, a solução passaria por recorrer a um número considerável de indivíduos para a definição dos dados de referência, realizando no final a sua média. No entanto, este processo exige um enorme volume de recursos (humanos e temporais) que não podem ser dispendidos nesta tarefa.

A geração dos dados de referência não é, ou não deve ser, da responsabilidade de cada investigador que pretenda avalizar uma nova técnica de segmentação e seguimento de objectos. Idealmente, a análise do desempenho dos novos algoritmos deve ser realizada por um mesmo conjunto de dados de referência, aberto e partilhado a toda a comunidade científica. Assim, no sentido de uniformizar a avaliação dos sistemas de seguimento de objectos, a *PETS* disponibilizou recentemente [PETS, 2001b; PETS, 2005] as trajectórias de referência, em formato XML. Outros projectos, como o [i-LIDS, 2006] e o [ETISEO, 2006] produziram dados de referência a partir de sequências de imagens que, contudo, apresentam restrições na sua utilização, i.e., são disponibilizados apenas aos grupos de investigação associados aos projectos ou são comercializados a preços elevados.

No que respeita à avaliação do desempenho das técnicas de segmentação de movimento, existem dois modos de realizar essa análise. O método mais elementar baseia-se nos objectos, requerendo apenas a definição das caixas delimitadoras dos objectos em movimento para todas as imagens da sequência de teste [Nascimento & Marques, 2004]. Um outro método, que possibilita uma avaliação mais precisa, requer um conjunto de dados de referência complexo, baseando-se no conjunto de pontos que definem a forma do objecto na imagem [Schlogl et al., 2004].

Por outro lado, a análise do desempenho das técnicas de seguimento de objectos faz apenas uso da informação fornecida pelas caixas delimitadoras dos objectos segmentados. Assim, quando se propõe avaliar um sistema que implementa a segmentação e o seguimento de objectos, dois esquemas de análise são possíveis:

- avaliação da técnica de segmentação de movimento baseada nos pontos que definem a forma dos objectos, complementada pela avaliação do método de seguimento baseada nos objectos (caixas delimitadoras); e

- avaliação de todo o sistema, i.e., técnica de segmentação e técnica de seguimento de objectos, baseada na informação dos objectos ao longo do tempo.

2.2. Classificação e Previsão de Comportamentos

Nos últimos anos assistiu-se a uma evolução significativa do aumento da capacidade de processamento. Suportadas nessa evolução tecnológica, técnicas de captura, processamento e análise de imagens digitalizadas experimentaram também um notório aperfeiçoamento. A prova do êxito destas novas tecnologias espelha-se no recente surgimento, sobre um amplo espectro de acção, de sistemas de vídeo-vigilância digitais para auxílio em tarefas de monitorização. Tais soluções possibilitam já efectuar a detecção automática de movimento, contagem de pessoas e detecção de multidões [Blueeyevideo, 2006], detecção de fumo e fogo [Fastcom, 2006], detecção de incidentes de tráfego [Citilog, 2006] e reconhecimento de matrículas de veículos motorizados [Survision, 2006].

Após os recentes progressos verificados na segmentação e seguimento de objectos em movimento, surge naturalmente o problema de interpretar e prever o comportamento dos objectos numa sequência de imagens. O reconhecimento de comportamentos pode ser visto como a classificação de sequências de dados, de tamanho variável, cujos atributos variam ao longo do tempo. Por outras palavras, pretende-se comparar cada sequência observada com um grupo de sequências de referência que representem comportamentos típicos. Tal comparação só será robusta e fiável, se a avaliação da similaridade entre sequências tolerar a existência de ruído nos dados, os problemas de escala e os de translação. Contudo, a avaliação da similaridade é condicionada em grande medida pelo modelo de representação dos dados, pelo que a sua escolha influenciará directamente a eficiência e a eficácia das técnicas propostas.

Embora os problemas da classificação e previsão de comportamentos tenham vindo a ser alvo de pesquisa durante mais de uma década, até ao momento as abordagens propostas não obtiveram um resultado que satisfizesse plenamente as exigências dos sistemas de vídeo-vigilância. Estas abordagens divergem, à partida, quanto à natureza dos modelos empregues na representação do comportamento, onde claramente são identificadas quatro categorias: **modelos não estocásticos**; **modelos discriminativos**; **modelos estocásticos descritivos**; e **modelos estocásticos generativos**.

Nesta secção são apresentados em primeiro lugar os trabalhos baseados em modelos não estocásticos, seguindo-se uma descrição dos trabalhos que fizeram uso de modelos

discriminativos. Por fim, descrevem-se alguns trabalhos baseados em abordagens de modelação estocástica descritiva e generativa.

2.2.1. Modelos Não Estocásticos

Os **modelos não estocásticos** foram os primeiros a ser utilizados na classificação e modelação de comportamentos. Este facto deve-se, muito provavelmente, à comodidade da sua aplicação, bem como à simplicidade dos conceitos teóricos subjacentes. Por força da sua natureza, os modelos não estocásticos impossibilitam qualquer tipo de aprendizagem automática. Assim, neste tipo de abordagem, os modelos necessitam forçosamente de ser definidos por um utilizador do sistema, o que implica o conhecimento prévio e exacto dos comportamentos que se pretendam observar.

Devido ao vasto leque de características que a modelação não estocástica permite considerar, as abordagens para a modelação de comportamentos divergem de forma significativa. Esta diversidade pode ser verificada na definição dos descritores dos comportamentos, na necessidade da existência ou não de conjuntos de treino, e ainda na contextualização do ambiente e objectos monitorizados.

Entre as técnicas mais populares de modelação não estocástica de comportamentos encontram-se: *Finite State Machines* (*FSMs*) [Gill, 1962], *Dynamic Time Warping* (*DTW*) [Sakoe & Chiba, 1978; Rabiner et al., 1978] e *Longest Common Subsequence* (*LCSS*) [Cormen et al., 1990].

Entre 1989 e 1993, no âmbito do projecto europeu *ESPIRIT II 2152*, ensaiou-se uma das primeiras abordagens à detecção automática de comportamentos, denominada por *Visual Inspection and Evaluation of Wide-area Scenes* (*VIEWS*) [Corrall, 1991; Howarth & Buxton, 1992]. O principal objectivo deste projecto consistia no desenvolvimento de técnicas que permitissem a automatização da vigilância de tráfego rodoviário através de câmaras de vídeo. A abordagem proposta nesse trabalho baseou-se na detecção de eventos predefinidos, utilizando informação espaço-temporal. Para tal, era necessário numa primeira fase identificar e classificar manualmente as diversas regiões monitorizadas pela câmara. A cada uma dessas regiões seria então atribuído, pelo utilizador, um determinado contexto (e.g. "*zona de entrada de veículos*", "*zona de saída*", "*zona de viragem à esquerda*", "*zona de viragem à direita*", "*rua pedestre*"). Estes dados seriam posteriormente inseridos numa base de dados de regiões.

A classificação dos eventos no sistema *VIEWS* executava-se através da comparação, para cada região predefinida, do tipo de movimento efectuado pelo objecto monitorizado com o contexto associado a essa região. Os autores deste trabalho defendiam que esta funcionalidade permitiria aferir o tipo de comportamento realizado pelo objecto. Como exemplo, postulavam que se um veículo se encontrasse parado sobre uma região classificada como de "*cedência de prioridade*", então o sistema indicaria que este estaria de facto a ceder passagem a um outro veículo.

O sistema *VIEWS* seguiu uma abordagem que manifestava sérias desvantagens. A necessidade e dependência da intervenção de um utilizador humano, para a definição de regiões observadas e atribuição dos respectivos contextos, colocavam em causa a eficácia e exactidão dos resultados obtidos pelo sistema. Deste modo, o desempenho do sistema dependeria de modo significativo do conhecimento que o utilizador possuísse acerca da área monitorizada. O facto de o comportamento de cada objecto ser aferido de um modo instantâneo, isto é, não ter em consideração o histórico do percurso realizado pelo objecto, consistia por si só um factor que colocava em causa a viabilidade desta abordagem. Além disso, tratava-se de um sistema estático, i.e., não se adaptava nem assimila novos comportamentos.

Com o projecto *Video Surveillance and Monitoring* (*VSAM*), pretendeu-se desenvolver tecnologia capaz de automatizar o processo da compreensão de acções, protagonizadas por humanos e veículos, observadas através de vídeo, para futuro uso em aplicações de vigilância urbana, bem como em campos de batalha. A abordagem proposta [Medioni et al., 2001] implicava a modelação do local a ser observado, de modo a possibilitar o relacionamento dos movimentos levados a cabo pelos objectos, com características relevantes do local. Por exemplo, certas velocidades e trajectórias em determinados contextos poderiam indicar um comportamento de ameaça.

A primeira fase na análise de comportamento, implementada pelo sistema *VSAM*, consistia em modelar o conjunto de actividades que se pretendesse reconhecer, através de dois tipos de contextos definidos *a priori*: contexto espacial; e contexto de missão. O contexto espacial correspondia a um mapa do local a observar, onde se definiam as estruturas espaciais, nomes simbólicos (e.g. estradas, zonas de pontos de controlo) e objectos estáticos de referência (e.g. ponto de controlo). Por seu lado, o contexto de missão continha os métodos específicos para o reconhecimento dos tipos de actividades definidas.

Após a contextualização do meio ambiente, o módulo de reconhecimento de actividades, que recebia para cada objecto dados referentes a oito atributos (altura, largura, velocidade, direcção do movimento e distância a um objecto de referência), analisava o tipo de actividade observada. Estas actividades poderiam ser compostas por vários estados, com transições entre estados activadas por eventos, como se de uma *Máquina de Estados Finitos* se tratasse.

Como se pode verificar, o processo de detecção de comportamentos desenvolvido no âmbito do projecto *VSAM* permitia ao utilizador definir, de uma forma simples, os tipos de comportamentos que se pretendessem detectar automaticamente. Esses comportamentos seriam descritos através de uma sequência de actividades (estados) de menor complexidade (e.g. "*parar*", "*virar à direita*", "*virar à esquerda*", "*acelerar*"). Porém, o modelo proposto implicava uma completa definição do contexto do local a monitorizar, bem como da correcta descrição das sequências de actividades que originassem comportamentos de risco. O sistema não possuía qualquer mecanismo que lhe permitisse descobrir novos tipos de comportamentos anormais, era pouco tolerante a ruído e não permitia prever nenhum tipo de comportamento.

Douglas Ayers e Mubarak Shah [2001] descreveram um sistema para detecção automática de acções humanas, observadas em sequências de vídeo, capturadas por uma câmara de vídeo a cores, em ambientes fechados (e.g. quartos ou salas de laboratórios). Como objectivo, os autores deste trabalho pretendiam detectar doze acções distintas, como por exemplo a entrada e saída de cena de um indivíduo, levantar e sentar-se, atender o telefone, usar um terminal de computador, etc. Para tal, era necessário que se fornecesse previamente ao sistema informações sobre os objectos relevantes no contexto das acções a observar, a disposição desses objectos na cena, bem como a definição das áreas de entrada e saída de indivíduos.

As acções eram modeladas por uma *Máquina de Estados Finitos* (*FSM*[4]). Cabia então ao utilizador definir os estados de acordo com as especificidades de cada ambiente monitorizado. Por outro lado, as transições entre estados eram verificadas com recurso a três componentes de baixo-nível (i.e. detecção de pele, seguimento de objectos e detecção de alteração na cena). A determinação de um estado implicava a informação acerca do

[4] Do inglês, *Finite State Machine.*

estado anterior bem como da validação das condições associadas a uma transição entre estados contíguos. Como exemplo, após a entrada de um indivíduo em cena, este seria seguido ao longo do tempo e, dependendo da sua interacção com o ambiente, poderia despoletar a transição de um estado "*em pé*" para o estado "*sentado*".

Os autores deste trabalho apontaram contudo uma limitação na utilização de *FSMs* para o reconhecimento de actividades humanas. Verificaram que o sucesso do sistema dependia de um modo muito significativo do conhecimento que o utilizador possuísse acerca das acções e do ambiente a monitorizar. Por exemplo, se uma área de entrada não fosse correctamente dimensionada, a detecção de entrada em cena de novos indivíduos poderia não ser correctamente identificada. Uma outra limitação do sistema está relacionada com a incapacidade de aprendizagem de novas actividades, bem como a detecção de entrada de novos objectos (não humanos) em cena.

Dan Buzan e seus colaboradores desenvolveram um sistema [Buzan, 2004; Buzan et al., 2004] que agrupava trajectórias similares em conjuntos hierárquicos, a partir de sequências temporais. As sequências de treino, compostas por informação da posição do objecto ao longo do tempo, eram agrupadas com o auxílio de um mecanismo que permitia calcular a similaridade entre trajectórias. Para o efeito, foi utilizada uma versão melhorada do algoritmo *Longest Common Subsequence* (*LCSS*) [Vlachos et al., 2002], originalmente apresentado em [Cormen et al., 1990]. Esta abordagem permitia a comparação de trajectórias adquiridas com diferentes períodos de amostragem, possibilitava a detecção de movimentos similares em diferentes regiões do espaço, bem como a análise de sequências com tamanhos distintos.

Os objectivos de Buzan não contemplavam, contudo, a previsão de comportamentos. Isto porque o método proposto não produzia modelos de padrões de comportamento. Desta forma, a classificação de novas sequências só seria possível se se mantivessem em memória todas as sequências observadas anteriormente. Este requisito tornaria o sistema impraticável em situações reais, quer pela elevada exigência de memória no armazenamento do histórico das sequências, quer pelo aumento da carga computacional à medida que o número de sequências fosse aumentando.

No trabalho apresentado em [Dee & Hogg, 2004], os autores propuseram uma abordagem para a previsão e classificação de comportamentos humanos, inspirada na *postura intencional*, defendida pelo filósofo Daniel Dennett [1987]. A *postura intencional* sugere que os seres

humanos e os animais são criaturas intencionais, i.e., os seus comportamentos são direccionados por objectivos, e a componente visível do comportamento pode frequentemente ser explicada pelos objectivos do agente.

Baseando-se na *postura intencional*, os autores construíram um modelo capaz de identificar padrões de comportamentos inexplicáveis, ou seja, comportamentos que não vão ao encontro do objectivo do agente observado. Esse modelo era composto por um conjunto de atributos que definiam um agente (nomeadamente coordenadas do centro de massa, etiqueta temporal e vector de velocidade do objecto), informação geográfica acerca da localização de obstáculos que afectassem o comportamento dos agentes, os seus objectivos e sub-objectivos.

Em cada amostragem e para cada objectivo, existiam então quatro possíveis relações entre cada agente e o seu objectivo. O objectivo poderia: estar directamente visível e no campo de visão do agente; estar directamente visível mas não se encontrar no campo de visão; não ser visível pelo agente, devido à existência de um obstáculo entre o agente e o seu objectivo; ser visível apenas por intermédio de um sub-objectivo.

A classificação de cada objectivo, como sendo consistente ou inconsistente com a trajectória observada, era então o passo seguinte. Para tal, os autores propuseram um diagrama de transição de estados, onde se definiam custos associados a cada transição. As transições em consonância com um determinado objectivo não implicavam qualquer custo. Por outro lado, transições associadas a movimentos divergentes do objectivo seriam penalizadas com um custo. Deste modo, e calculando o custo total para cada objectivo da cena, o objectivo com o menor custo total era o que melhor explicaria o comportamento do agente.

Os resultados experimentais do trabalho desenvolvido por Dee e Hogg indicaram uma variação acentuada da correlação[5], na detecção de comportamentos inexplicáveis, entre o sistema proposto e um conjunto de supervisores humanos. Os resultados permitiram ainda verificar que a precisão na classificação de comportamentos variava de forma significativa com o tipo de cena observada. Por exemplo, num cenário de parque de estacionamento

[5] Na estatística, o coeficiente de correlação de Pearson mede a relação linear entre duas variáveis, e toma valores ente -1 (correlação perfeitamente negativa) e 1 (correlação perfeitamente positiva).

obteve-se uma baixa correlação (0.35), enquanto que num cenário de um átrio de um edifício se obteve uma correlação mais elevada (0.639).

Em [Pierobon et al., 2005], os autores propuseram um método para *clustering* de actividades humanas, baseado numa representação tridimensional do corpo em termos de coordenadas volumétricas. Essa representação era conseguida, para cada nova amostra, através da junção de imagens provenientes de oito câmaras de vídeo sincronizadas e em diferentes pontos de vista que, por meio de uma técnica de intersecção volumétrica [Yue et al., 2003], possibilitava a construção de um modelo tridimensional do actor na cena monitorizada.

O princípio empregue por Pierobon no seu trabalho, considerava que uma actividade humana podia ser descrita por uma sequência de posturas, adaptada de modo a ser independente da posição, tamanho, escala e proporções corporais de um indivíduo. Assim, um vector em que se identificavam mil características era utilizado para caracterizar a postura corporal num ambiente estático [Cohen & Li, 2003]. Uma qualquer actividade seria então definida por uma sequência de vectores de características ao longo do tempo.

O reconhecimento de actividades realizava-se pela aplicação de uma medida de distância entre sequências, em que se comparava cada sequência de teste com as sequências de referência. Para tal, os autores recorreram à técnica *Dynamic Time Warping* (*DTW*) [Sakoe & Chiba, 1978; Rabiner et al., 1978], uma vez que esta possibilitava a comparação de sequências de dimensões diferentes, suportando também possíveis distorções não lineares. O caminho óptimo da matriz de *warping* era obtido pelo cálculo do custo mínimo total, de acordo com o princípio de optimização de Bellman [Theodoridis & Koutroumbas, 1999].

O método proposto por Pierobon foi testado na identificação de três actividades simples: "*apontar para*", "*baixar-se*" e "*pontapear*". Apesar de os resultados apontarem a aptidão do método para a classificação de actividades, este exibe importantes limitações. O modelo proposto requer elevados recursos computacionais, quer na construção do modelo tridimensional do corpo humano, quer na identificação de actividades. Uma outra limitação bastante conhecida da técnica de *DTW* reside no facto da complexidade computacional estar directamente dependente das dimensões das sequências a comparar, bem como do número de sequências de referência predefinidas [Keogh & Ratanamahatana, 2005].

2.2.2. Modelos Discriminativos

Na classificação com **modelos discriminativos**, também designados por modelos de classificação, o objectivo consiste em mapear, de uma forma directa, um vector de entrada (contendo dados monitorizados, como por exemplo a área e a velocidade média de um objecto) num símbolo (classe), pertencente ao conjunto de símbolos possíveis. Por outras palavras, dado um espaço multi-dimensional definido pelos vectores de entrada, pretende-se definir limites ou superfícies de decisão, entre regiões que definam símbolos distintos que, no âmbito deste trabalho, se referem a classes de comportamento.

Em problemas de classificação reais, contudo, as classes não se encontram perfeitamente separadas por limites de decisão. Verifica-se frequentemente a sobreposição de elementos de duas ou mais classes em determinadas regiões do espaço. Assim, ao invés de utilizar superfícies de decisão, recorre-se então a uma função que maximize alguma medida de separação entre classes. A essas funções dá-se o nome de funções discriminantes.

De entre os modelos discriminativos mais relevantes encontram-se as seguintes técnicas: *Redes Neuronais* (*NNs*) [Davalo, 1988]; *K-Nearest Neighbour* (*K-NN*) [Cover & Hart, 1967]; *Self-Organizing Map* (*SOM*) [Kohonen, 1997]; *Support Vector Machines* (*SVMs*) [Burges, 1998]; *Maximum Entropy Markov Model* (*MEMM*) [McCallum et al., 2000]; e *Conditional Random Fields* (*CRF*) [Lafferty et al., 2001].

Neil Johnson e David Hogg publicaram um trabalho [1996] onde, utilizando uma abordagem baseada num modelo estatístico, conferiram ao sistema a capacidade para aprender, de uma forma não supervisionada, as trajectórias típicas realizadas por pedestres observadas ao longo de sequências de imagens. O modelo de aprendizagem proposto por estes dois investigadores consistia na combinação da técnica *Vector Quantization* (*VQ*) [Linde et al., 1980], implementada por *Redes Neuronais Competitivas* (*CNN*[6]) [Rumelhart & Zipser, 1986], com um tipo de neurónio de memória de curto prazo.

A rede neuronal proposta era constituída por quatro neurónios de entrada, alimentados por dados de treino com atributos referentes às coordenadas do centro de massa do objecto e do deslocamento (horizontal e vertical) sofrido desde a posição anterior. Como resultado, a rede neuronal geraria mil protótipos de estado, i.e. vectores que caracterizavam o tipo de

[6] Do inglês, *Competitive Neural Network*.

deslocamento dos objectos no espaço bidimensional. O passo seguinte consistia em interligar cada neurónio de protótipo de estado com um neurónio de memória de curto prazo. Estes mil neurónios estariam, por sua vez, conectados a uma rede neuronal com uma camada de saída de quinhentos neurónios que definiam os protótipos das trajectórias.

O sistema proposto possibilitava dois tipos de reconhecimento de eventos. Um primeiro tipo permitia o reconhecimento de eventos simples, através da atribuição de semânticas aos protótipos de estado, por parte do utilizador. Deste modo, a sinalização dos eventos efectuava-se ao longo da trajectória do objecto, de acordo com os protótipos de estado atribuídos durante o deslocamento. Um outro tipo de eventos, de maior complexidade, era detectado nos casos em que uma determinada trajectória se desviasse de forma significativa dos protótipos construídos a partir dos dados de treino.

Apesar de permitir efectuar, de uma forma autónoma, a aprendizagem de trajectórias frequentes, o sistema proposto por Johnson e Hogg apresentava diversas limitações. A utilização de uma rede neuronal implicava uma dependência directa entre a precisão e exactidão dos resultados da detecção de eventos e a dimensão da rede, nomeadamente do número de neurónios de saída. O sistema não possuía a capacidade de se adaptar ao longo do tempo, i.e. uma vez treinado, este apenas se limitava a diferenciar trajectórias comuns de trajectórias anormais. A estas limitações acrescia o facto de este tipo de abordagem não permitir efectuar nenhum tipo de previsão de ocorrência de eventos anormais.

Em [Zhong & Shi, 2003; Zhong et al., 2004] foi proposto um algoritmo que efectua a classificação e detecção de actividades humanas, sem necessidade de supervisão. A abordagem apresentada não recorre a dados resultantes do seguimento de objectos. Ao invés, utiliza apenas características visuais do movimento global ocorrido em cada imagem, nomeadamente histogramas de movimento e de cor.

O método proposto obriga a um pré-processamento da sequência de vídeo, previamente adquirida ao longo de várias horas. Assim, numa primeira fase, procedia-se à redução da quantidade de imagens a analisar, através da remoção de imagens que não possuíssem qualquer objecto em movimento. A preparação da sequência prosseguia com uma nova redução, através da amostragem de uma imagem em cada dez. Todavia, este procedimento poderia comprometer informação relevante sobre certas actividades.

Cada imagem da sequência era posteriormente representada por um histograma de movimento e um histograma de cor das regiões classificadas como movimento. Estes histogramas eram vistos como um vector característico. Aproveitando o facto de os vectores serem altamente redundantes, uma compressão por *VQ* era então utilizada de modo a reduzir a dimensão dos dados.

Seguidamente, realizava-se a decomposição da sequência de imagens, resultantes do processo anterior, em pequenos segmentos de vídeo sobrepostos, e.g. segmentos de quatro segundos (0s a 4s, 2s a 6s, 4s a 8s). Contudo, a escolha do tamanho dos segmentos requeria um conhecimento da duração dos eventos que se pretendiam classificar, tendo um papel relevante no sucesso desta abordagem. Para eventos de curta duração, segmentos de quatro segundos podiam abranger com sucesso actividades como um indivíduo a deixar cair algo, ou a pegar um objecto da cena. Contudo, para actividades de longa duração, o período deveria ser consideravelmente aumentado, de modo a capturar as características relevantes do evento. Esta particularidade do método proposto implicava que todas as actividades a identificar tivessem períodos de execução similares.

Partindo do pressuposto que numa longa sequência de imagens existe sempre um número reduzido de acções, se comparado com o número de segmentos de vídeo, então estes segmentos deveriam ser agrupados em conjuntos que representem eventos distintos. Esse agrupamento era executado através da medição da similaridade dos diferentes segmentos. Para tal, eram considerados os protótipos dos vectores característicos, resultantes da compressão por *VQ*, associados a cada segmento.

Por fim, a classificação de eventos era realizada para cada novo segmento de vídeo. Para tal, o segmento era analisado de modo a encontrar os protótipos dos vectores característicos. A classificação seria então efectuada por meio da aplicação de um algoritmo *K-Nearest Neighbour* (*K-NN*) [Cover & Hart, 1967]. Eventos anormais seriam detectados no caso de lhe serem atribuídos conjuntos de segmentos isolados, i.e. que ocorressem com reduzida frequência.

Em [Khalid & Naftel, 2005], foi proposta uma nova técnica para classificação de trajectórias de objectos, baseada numa rede neuronal *Self-Organizing Map* (*SOM*) [Kohonen, 1997], com o objectivo de descobrir similaridades entre trajectórias, de uma forma não supervisionada. Nesse trabalho, as trajectórias eram representadas por duas sequências temporais unidimensionais, i.e., com as componentes horizontal e vertical da sequência de

coordenadas do centro de massa de um objecto, caracterizadas separadamente em função do tempo.

De modo a obter uma redução de dimensão, procedia-se então ao cálculo dos coeficientes de *Fourier*, aplicando a *Transformada Rápida de Fourier* (*FFT*[7]) [Cooley & Tukey, 1965] sobre as sequências temporais. O coeficiente constante era removido, de forma a normalizar os dados de cada sequência, sendo que os primeiros três coeficientes constituíam o vector de entrada da rede *SOM*. A topologia desta rede consistia numa camada de neurónios de entrada, alimentada pelos coeficientes de *Fourier*, conectados directamente a uma camada unidimensional de neurónios de saída. Cada neurónio de entrada encontrava-se ligado a todos os neurónios de saída, com a conexão representada por um vector peso.

O número de neurónios de saída definia a quantidade de padrões de trajectórias distintas que se pretendiam descobrir. Contudo, este número é definido manualmente pelo utilizador, sendo que, após efectuar o treino da rede, padrões similares são aglomerados em *clusters*. No entanto, o número final de *clusters* de trajectórias é definido empiricamente, o que implica um conhecimento prévio do número aproximado de trajectórias a observar.

As trajectórias anormais são detectadas através da análise da distância de *Mahalanobis* entre uma nova trajectória e o centro do *cluster* mais próximo. Assim, o sistema permite distinguir entre trajectórias normais (trajectórias previamente adquiridas na fase de treino) de trajectórias anormais (novas trajectórias que não tivessem sido observadas anteriormente).

Debruçando-se sobre o reconhecimento de acções, protagonizadas por jogadores de ténis durante um encontro, Guangyu Zhu e seus colaboradores em [2006] apresentaram uma nova abordagem na classificação de actividades. Para tal, propuseram a descrição do movimento de cada atleta em histogramas de fluxo óptico e a aplicação da técnica *Support Vector Machines* (*SVMs*) [Burges, 1998] na classificação de duas actividades distintas: "*bater a bola à esquerda*" e "*bater a bola à direita*".

Explorando as particularidades das acções a monitorizar, os autores sugeriram um método para descrição do movimento baseado em quatro histogramas de fluxo óptico. Assim, numa primeira fase, realizava-se o cálculo do fluxo óptico pelo algoritmo de *Horn-Schunck* [Horn & Schunck, 1981] numa região de interesse que abrangia o jogador e uma área

[7] Do Inglês, *Fast Fourier Transform*.

predeterminada à sua volta. Seguidamente, essa região era dividida verticalmente em três sub-regiões, que abrangiam as áreas à esquerda do jogador, a área ocupada pelo atleta (centro), e a área à sua direita. Após a eliminação de ruído sobre os resultados obtidos pelo cálculo do fluxo óptico, procedia-se então à construção dos quatro histogramas. Dois histogramas identificavam as projecções dos vectores de fluxo óptico sobre os eixos horizontal e vertical da área à esquerda do atleta, sendo que os restantes dois histogramas apresentavam as projecções dos vectores de fluxo óptico nos dois eixos, da área à direita do atleta.

Dado que a *SVM* é uma técnica de aprendizagem supervisionada, os autores definiram um conjunto de imagens de teste, devidamente classificadas como pertencentes a uma de três classes: "*bater a bola à esquerda*", "*bater a bola à direita*" e "*sem batimento*". Após o treino da *SVM* e da resultante obtenção dos vectores de suporte, o classificador demonstrou uma elevada taxa de acerto (aproximadamente 87%) no reconhecimento das duas actividades, em imagens isoladas.

O reconhecimento de actividades em sequências de vídeo foi realizado com o recurso a este classificador, através de um sistema de votação. Assim, dada uma sequência de imagens, conjuntos de vinte e cinco imagens adjacentes, que antecediam o batimento de uma bola pelo atleta, eram processadas pelo classificador *SVM*. A detecção do batimento da bola realizava-se com o auxílio de informação áudio, utilizando uma técnica introduzida por Xu [Xu et al., 2003]. Para cada imagem analisada, o classificador atribuía a respectiva classe, sendo que, após o processamento de cada conjunto de imagens, a actividade detectada era definida pela classe que obtivesse maior frequência.

2.2.3. Modelos Estocásticos Descritivos

Os **modelos estocásticos descritivos**, pela sua natureza, permitem realizar uma aprendizagem dos comportamentos, sem necessidade de supervisão. Estes modelos, que não possibilitam qualquer tipo de previsão, focam-se na estrutura intrínseca dos dados e na descoberta de relações. O seu objectivo consiste em criar um modelo que descreva as principais características dos dados, ou seja, pretende-se essencialmente sumariar os dados. Essa descrição pode ser realizada por modelos de distribuição de probabilidades dos dados, por seccionamento de um espaço multi-dimensional em diversos grupos (*clusters*), e pela descrição de relacionamentos entre variáveis (modelação de dependências).

São exemplos de modelos estocásticos descritivos as técnicas: *K-Means* [MacQueen, 1967]; *Expectation-Maximization* (*EM*) [Dempster et al., 1977]; *Numeric Iterative Hierarchical Cluster* (*NIHC*) [Wallace, 1989]; e *O-Cluster* [Milenova & Campos, 2002].

Em [Grimson et al., 1998], os autores descreveram um sistema capaz de efectuar a aprendizagem de padrões de actividades comuns, protagonizadas por diferentes objectos. Para tal, representaram as actividades num espaço de seis dimensões, i.e., coordenadas bidimensionais da posição do centro-de-massa do objecto, a sua velocidade (horizontal e vertical), o tamanho do objecto e a relação entre a sua altura e largura. O sistema proposto permite ainda realizar a detecção de actividades não usuais, através de uma análise comparativa com os padrões previamente descobertos.

Grimson e a sua equipa de investigação, ensaiaram duas abordagens na classificação de actividades. O primeiro método consistia no agrupamento das observações no espaço hexadimensional, pelo uso de um algoritmo de *NIHC* [Wallace, 1989]. Mediante o uso deste algoritmo, os dados eram agrupados através de um processo iterativo de *clustering* hierárquico. Como resultado, os dados seriam apresentados numa árvore binária, que definia a estrutura dos vários *clusters*.

O segundo método, explorado em [Grimson et al., 1998], baseava-se na acumulação de estatísticas de co-ocurrência num espaço hexadimensional quantizado. Nesta abordagem, as trajectórias eram representadas por uma sequência de estados, obtidos por um algoritmo de *K-Means* aplicado sobre os dados. Uma matriz de co-ocurrência, em que se assumia que todas as sequências tinham o mesmo tamanho, era então elaborada. Cada elemento dessa matriz definia a probabilidade de dois estados ocorrerem numa mesma sequência.

Dimitrios Makris e Tim Ellis têm vindo a desenvolver um paradigma de aprendizagem, não supervisionado, de modelos de semântica do ambiente monitorizado. O objectivo destes investigadores é o de identificar regiões, no plano da imagem, onde se verifiquem certos tipos de movimentos ou actividades. Para tal, utilizam informação espacial proveniente do processo de seguimento de objectos.

Numa primeira fase [Makris & Ellis, 2003], cinco tipos de áreas eram classificadas, nomeadamente "*zonas de entrada*", "*zonas de saída*", "*caminhos*", "*rotas*" e "*junções*". As *zonas de entrada* definiam áreas onde os objectos entravam em cena. Similarmente, as *zonas de saída* indicavam as regiões onde os objectos abandonavam a cena. As *junções*, por seu lado,

identificavam áreas onde dois ou mais caminhos se interceptavam, enquanto que os *caminhos* definiam segmentos entre "*zonas de entrada*"/"*saída*" e "*junções*". Por outro lado, as *rotas* consistiam no trajecto que um objecto percorresse, desde a entrada até à sua saída, passando por vários *caminhos* e *junções*. Numa segunda fase [Makris & Ellis, 2005], um outro tipo de área, denominada por "*zona de paragem*", foi adicionada ao modelo.

As "*zonas de entrada*", "*saída*" e de "*paragem*" são definidas por *Modelos Gaussianos* (*GMs*[8]). O recurso a *GMs* possibilita a definição espacial das zonas, enquanto proporciona uma base para análise probabilística. Por conseguinte, procede-se à criação de três tipos de conjuntos de dados. Um conjunto com pontos que definem as coordenadas de entrada dos objectos em cena, um outro conjunto que identifica pontos com as coordenadas da última posição de cada objecto em cena e, por fim, o conjunto que reúne pontos onde se tivesse verificado a paragem de objectos.

Cada um destes conjuntos era então submetido a um algoritmo de *EM*, de modo a identificar diferentes distribuições nos dados. O algoritmo *EM* empregue neste trabalho necessitava da definição prévia do número de classes a observar, o que implicava que o utilizador do sistema tivesse um conhecimento prévio de como os dados se distribuíam no espaço. Contudo, este problema foi minorado através de uma função de ordenamento que consistia na razão da percentagem de pontos pertencentes à classe, pela área da elipse gaussiana. As classes que obtivessem um resultado, da função de ordenamento, superior a um determinado valor eram consideradas como válidas, enquanto as restantes eram tratadas como ruído e, por conseguinte, eliminadas.

As *rotas*, por seu lado, não podiam ser representadas através de *GMs*, necessitando então de um modelo mais descritivo. Para tal, foi proposto em [Makris & Ellis, 2001; 2002] um modelo espacial capaz de representar o eixo principal, a largura, o uso e a sua distribuição ao longo da rota. Nestes dois trabalhos, foi ainda apresentado um algoritmo que permitia a aprendizagem de modelos de *rotas* a partir de um conjunto de trajectórias, sem qualquer tipo de inicialização. Depois de definidas as *rotas*, as *junções* podiam ser facilmente derivadas do conjunto de *rotas*, pela observação de áreas sobrepostas de duas ou mais *rotas*.

Apesar da aplicação óbvia deste modelo semântico ser a anotação de sequências de imagens de vídeo, por exemplo, na descrição textual de uma actividade, existem outros domínios

[8] Do inglês, *Gaussian Model* (*GM*).

onde o método pode ser aplicado. Nomeadamente, no cálculo da zona provável de saída, ou ainda, dado que as *rotas* aprendidas identificam os padrões típicos de comportamento, na identificação de comportamentos anormais sempre que um trajecto se desvie das rotas estabelecidas.

2.2.4. Modelos Estocásticos Generativos

Os **modelos estocásticos generativos** permitem criar um modelo estatístico de comportamento futuro. Neste tipo de modelação os resultados, ou comportamentos futuros, são influenciados por um conjunto de factores variáveis. Na previsão de comportamentos em vídeo-vigilância, por exemplo, a velocidade, a cor, e a posição de um objecto ao longo do tempo, podem possibilitar a previsão da ocorrência de eventos anormais.

O objectivo da abordagem generativa consiste em definir um modelo específico que interprete o modo como as observações (dados) são geradas. Para tal, assume-se a existência de um conjunto de variáveis ocultas, organizadas de forma desconhecida, responsáveis pela geração dos dados observados. O desafio da aprendizagem generativa consiste então na identificação das variáveis ocultas e do modo de como estas se relacionam.

O sucesso de um modelo generativo depende da capacidade deste adquirir a estrutura do fenómeno subjacente às observações. Várias técnicas foram propostas com este intuito, sendo as mais populares a *Principal Components Analysis* (*PCA*) [Jolliffe, 1986], *Modelos de Misturas de Gaussianos* (*GMM*[9]) [McLachlan & Basford, 1988], *Hidden Markov Models* (*HMM*) [Rabiner, 1989], e *Redes Bayesianas* (*BN*) [Duda et al., 2001].

Os modelos generativos foram utilizados para a detecção de comportamentos, numa segunda fase do projecto *VIEWS*. Em [Buxton & Gong, 1995], os autores propuseram uma *Rede Dinâmica de Bayes* (*DBN*[10]) associada a um motor de inferência, a operar em sequências de imagens de tráfego de uma auto-estrada, com o intuito de gerar conceitos de alto nível, como "*veículo mudando de faixa*" ou "*veículo parado*". Conceptualmente, os autores deste trabalho pretendiam modelar, de uma forma incompleta e consentindo um elevado

[9] Do inglês, *Gaussian Mixture Model.*

[10] Do ingles, *Dynamic Bayesian Network.*

grau de incerteza, um sistema capaz de recolher acções a partir de sequências de imagens, para suportar interpretações simples e predefinidas de comportamentos observados na área monitorizada. Esta nova abordagem adoptava, contudo, a restrição da proposta inicial do projecto, em que era necessário utilizar um modelo pré-calibrado da câmara, juntamente com um modelo geométrico da zona a monitorizar.

A *DBN* proposta em [Buxton & Gong, 1995], formada por sub-grafos temporalmente separados, definia na sua estrutura dois tipos de nós que, por sua vez, apresentavam dois tipos de relações: espacial e temporal. A inferência *bayesiana* implica que cada nó, consistindo num conjunto exaustivo de estados mutuamente exclusivos, deve ter um conjunto de probabilidades condicionais *a priori*. A limitada variedade de estados, bem como a qualidade e a forma de como os dados eram obtidos, constituíam factores que influenciavam directamente e de modo negativo o resultado do classificador.

Apesar de apresentar relevantes limitações, a utilização de *DBN* continuou a ser explorada por Richard Howarth e Hilary Buxton no projecto *HIVIS-WATCHER* [Buxton & Howarth, 1996; Buxton, 1997; Howarth & Buxton, 1998], não se verificando, contudo, avanços significativos em relação à proposta inicial.

Uma outra abordagem na detecção de comportamentos foi ensaiada por Yaser Yacoob e Michael Black [Yacoob & Black, 1999]. Os autores propuseram um mecanismo para reconhecimento de actividades humanas, mais concretamente de actividades atómicas. Este tipo de actividades caracteriza-se, ao nível da sua representação, através de medidas de um conjunto de atributos durante uma janela temporal finita e de curta duração, e.g. o movimento das pernas de um indivíduo a caminhar, durante um ciclo. O modelo de reconhecimento de actividades atómicas, baseado em *Principal Components Analysis* (*PCA*) e transformações lineares, implicava a definição prévia do conjunto de actividades a monitorizar (caminhar, chutar e marchar), de forma a reunir dados de treino, de diferentes partes do corpo humano, que descrevessem essas actividades.

Através da utilização da técnica *PCA* era possível obter um modelo de actividades de menor dimensão, i.e., onde seriam retirados os componentes menos significativos, reduzindo assim a quantidade de informação. O reconhecimento das actividades seria realizado por meio da comparação dos componentes principais de uma actividade observada, com os modelos anteriormente aprendidos.

O emprego da técnica *PCA* em reconhecimento de padrões, sobre dados com um elevado número de variáveis, mostrou a sua viabilidade no trabalho desenvolvido por Yacoob e Black. Todavia, esta técnica não permitia efectuar uma aprendizagem *on-line* e apenas possibilitava o reconhecimento de actividades predefinidas e simples, o que seria manifestamente insuficiente para uma funcionalidade mais complexa, como é o caso de detecção de comportamentos anormais.

Um outro projecto, denominado por *VIGILANT* [Remagnino & Jones, 2001], cujo objectivo consistia em desenvolver um sistema para descrição de eventos observados num ambiente típico de um parque de estacionamento automóvel, apresentava uma nova abordagem à problemática da detecção de eventos. Nesse sistema, o processo de seguimento de objectos, que operava em tempo-real, era complementado por um processo *off-line*, para classificação de eventos, executado apenas em períodos de reduzida actividade na cena. Este processo gerava uma classificação do tipo de objecto (humano ou veículo), histórico de cor, e uma descrição semântica do evento associado ao objecto monitorizado.

A orientação deste trabalho apontava sobretudo para a produção de anotações das características dos objectos e seus eventos, que seriam arquivadas numa base de dados, de modo a facilitar o posterior processo de consulta de gravações de vídeo. Deste modo, seria possível ao utilizador pesquisar e consultar eventos, previamente observados, através da formulação de um conjunto de perguntas, com vocabulário controlado.

A classificação de comportamentos no projecto *VIGILANT* é implementada através de *HMMs*. Cada tipo de comportamento é definido por um modelo, cujo número de estados é previamente definido, e em que: os estados representam regiões da imagem; a probabilidade *a priori* especifica a probabilidade de um evento começar numa determinada região; e as probabilidades de transição são definidas pela probabilidade de progressão entre estados.

Esta abordagem implica que, numa primeira fase, fossem construídos os modelos dos comportamentos que se pretendesse reconhecer. Para proceder à sua construção, é necessário reunir um conjunto de dados de treino que descrevam os comportamentos alvo de interesse. Após a sua construção, estes modelos, que são utilizados na posterior execução das tarefas de classificação de eventos, permanecem imutáveis ao longo do tempo. Ou seja, o sistema não tem capacidade para assimilar novos tipos de comportamentos.

Os testes realizados pelos autores deste trabalho sobre o classificador de eventos, para modelos de cinco a vinte estados, revelaram uma percentagem de eventos correctamente detectados variável entre 55% e 75%. Ou seja, verificou-se que com apenas cinco estados a classificação exibia características praticamente aleatórias (55%). A percentagem de acerto melhorava, contudo, com o aumento do número de estados. Os testes apresentados não avaliavam a taxa de falsos positivos do classificador e apenas quatro tipos de eventos simples foram alvo de avaliação (entrada de veículo, saída de veículo, entrada de pessoa e saída de pessoa).

Rocío León e Luis Sucar [2002] apresentaram um modelo de reconhecimento de diferentes actividades, executadas continuamente, a diversas velocidades e por diferentes pessoas. As actividades, que neste trabalho estavam reduzidas apenas a "*acenar*" e "*mover para a direita*" a mão direita de um indivíduo, eram modeladas pela trajectória global, através da direcção do movimento da mão entre imagens sucessivas.

Cada uma destas actividades era modelada por uma *BN*, cuja raiz representava a actividade a reconhecer e em que os nós definiam nove possíveis direcções de deslocamento. O problema de determinar o número de imagens necessárias para representar uma actividade foi superado através do uso da análise de *Fourier*. Com efeito, a *Transformada Discreta de Fourier* (*DFT*[11]) foi utilizada para determinar o período de tempo necessário para representar uma actividade.

Depois de construídos os modelos de actividade, a cada sequência de trinta imagens era efectuada uma análise de *Fourier*, para determinar o período de repetição de um gesto. Seguidamente, o número de imagens nesse período era normalizado para dez imagens. Nessas imagens, os atributos que descreviam a deslocação da mão direita eram então aplicados aos dois modelos, de modo a calcular qual a actividade que obtinha maior probabilidade, i.e. aferir qual a actividade observada.

Em 2002, Neil Johnson e David Hogg propuseram um novo método de síntese de comportamentos que empregava um *Modelo de Mistura de Gaussianos* (*GMM*) com dez componentes [Johnson & Hogg, 2002]. Neste método, o estado do sistema num determinado momento era definido pelas coordenadas bidimensionais do centro de massa do objecto e por uma representação de comprimento fixo (*t*, *t-1* e *t-2*, em que *t* representa

[11] Do inglês, *Discrete Fourier Transform.*

uma etiqueta temporal) do histórico dos estados. Por conseguinte, o sistema era definido por um conjunto de funções densidade de probabilidade num espaço de oito dimensões.

Os parâmetros que definiam as dez funções densidade de probabilidade do *GMM* eram obtidos, numa fase de treino, através da aplicação de uma estimativa por *VQ* seguida pela aplicação de um algoritmo *EM*. Deste modo, era possível reduzir significativamente o número de iterações levadas a cabo pelo algoritmo *EM*. Após o treino e definidos então os parâmetros do *GMM*, a síntese do comportamento era obtida por meio da previsão do deslocamento, realizada pela selecção do deslocamento que maximizasse a probabilidade condicional desse deslocamento, para um determinado histórico de estados.

Os autores do método demonstraram ainda ser possível efectuar com sucesso a codificação de comportamentos não lineares. Porém, os testes experimentais careciam de uma avaliação mais profunda, uma vez que apenas foram utilizadas quinze trajectórias. Um dos inconvenientes deste método deve-se à necessidade de efectuar uma definição prévia e estática do número de componentes do *GMM*, sem se proceder a uma análise prévia da distribuição dos dados no espaço. Neste trabalho, apesar de se abordar a síntese de comportamentos, nada é referido sobre a possível utilização das actividades sintetizadas para a classificação de comportamentos.

Em [Hongeng et al., 2004], foi apresentado um novo método para o reconhecimento de actividades humanas. Neste trabalho, uma actividade pode ser composta por eventos (simples ou complexos) executados por um indivíduo, ou formada por vários eventos que modelem interacções entre indivíduos (eventos múltiplos). A modelação dos eventos, a partir de características da forma e trajectória de um indivíduo, é realizada através de uma representação hierárquica de actividades [Medioni et al., 2001], onde os eventos são organizados em níveis de abstracção, proporcionando flexibilidade e modularidade na modulação das actividades.

Os eventos simples são calculados, em cada instante, recorrendo a *BNs* construídas a partir das propriedades do indivíduo (forma do objecto e sua trajectória). As *BNs* permitem especificar, de uma forma directa, relações causais entre as propriedades dos indivíduos e eventos simples. Contudo, cabe ao utilizador do sistema definir a estrutura da *BN* que modela cada evento, ou seja, definir os sub-eventos e as relações causais entre eles. Como exemplo, o evento simples "*o indivíduo aproxima-se da pessoa de referência*" seria descrito pela

ocorrência de três sub-eventos: "*estar perto da pessoa de referência*", "*deslocar-se na direcção de*" e "*abrandar*".

A modelação de uma sequência temporal, ordenada, de eventos simples, levados a cabo por um indivíduo, é composta por eventos complexos. A modelação deste tipo de eventos é assegurada através de *Autómatos de Estados Finitos* [Hongeng et al., 2000], que permitem uma descrição natural dos eventos. Por exemplo, o evento complexo "*conversar*" deverá ser visto como uma ocorrência linear de dois eventos simples consecutivos: "*o indivíduo aproxima-se da pessoa de referência*", seguido do evento "*pára junto à pessoa de referência*".

Eventos múltiplos modelam dois ou mais eventos, simples ou complexos, com uma relação lógica e temporal entre eles. Ou seja, descrevem interacções entre diferentes indivíduos. Este tipo de evento é representado por um grafo de eventos, onde as ligações entre os nós descrevem as relações temporais entre os sub-eventos de um evento múltiplo, e.g. "*evento A deve ocorrer antes do evento B*" ou "*ocorre o evento A ou evento B*".

Somboon Hongeng e seus colaboradores [2004] apontaram algumas fragilidades que podem afectar o desempenho do algoritmo de reconhecimento. Nomeadamente, a possibilidade das *BNs* serem afectadas por ruído e erros provenientes dos processos de segmentação e seguimento de objectos em movimento. Durações variáveis de sub-eventos podem também comprometer o sucesso da detecção de eventos complexos levada a cabo pelos *Autómatos de Estado-finito*. Por fim, os grafos de eventos, utilizados para a modelação de eventos múltiplos, são sensíveis à variação do sincronismo de relações entre eventos, devido a diferentes estilos de execução das acções.

Um outro sistema capaz de reconhecer actividades humanas complexas, observadas em ambientes externos, foi apresentado em [Leo et al., 2004; 2004b]. A abordagem proposta explora as variações das posturas dos indivíduos ao longo do tempo e destina-se ao reconhecimento de quatro actividades distintas ("*caminhar*", "*sondar um local*", "*molhar o solo*" e "*apanhar um objecto*"), típicas de um local de escavações arqueológicas.

O sistema processa as máscaras binárias, resultantes da segmentação de objectos em movimento, de modo a efectuar uma estimativa da postura do indivíduo, de entre três tipos predefinidos: "*em pé*", "*abaixado*" e "*inclinado*". Utilizando uma medida de proximidade entre posturas, calculada através da distância de *Manhattan* dos histogramas das projecções horizontal e vertical da máscara de movimento do indivíduo, um algoritmo de *clustering* não

supervisionado, denominado por *Basic Competitive Learning Scheme* (*BCLS*) [Theodoridis & Koutroumbas, 1999] agrupa, numa primeira fase, as imagens de treino. Posteriormente, os protótipos de posturas são utilizados pelo algoritmo *BCLS* na classificação de novas posturas observadas.

O reconhecimento de comportamentos efectua-se com o recurso a *HMMs*, inteiramente conectados, i.e., com cada estado de um modelo ligado a todos os estados. A opção por *HMMs* inteiramente conectados possibilita a modelação de processos estatísticos complexos. Contudo, a sua operação evidencia um maior grau de dificuldade, devido ao elevado número de parâmetros a definir.

Na abordagem proposta, cada actividade está associada a um *HMM* e o número de posturas determina o número de símbolos (*codebook*) do modelo. No entanto, o número de estados não pode ser definido previamente, tendo que ser testado para cada tipo de actividade, de modo a determinar o seu valor óptimo.

Uma relevante limitação deste método de reconhecimento de comportamentos, reside na obrigação de que as sequências fornecidas aos *HMMs*, tanto na fase de treino como na fase de teste, possuírem tamanho fixo, e da especificação desse tamanho requerer uma prévia avaliação experimental. Os autores verificaram ainda que, no método proposto, sequências de curta duração não caracterizavam satisfatoriamente qualquer actividade relevante e que, por outro lado, longas sequências não possibilitavam a generalização de diferentes execuções de uma mesma actividade.

Em [Arsic et al., 2005], no âmbito do projecto *Security of Aircraft in the Future European Environment* (*SAFEE*), foi proposto um sistema para a detecção automática de comportamentos anormais de passageiros de veículos de transporte público (aviões e comboios), monitorizados por uma câmara de vídeo fixa. Em virtude das características físicas dos transportes, os autores deste trabalho pretenderam detectar comportamentos através da observação de acções realizadas pela parte superior do corpo dos passageiros, nomeadamente o tronco e a cabeça.

Arsic e seus colaboradores postularam que determinados comportamentos (e.g. nervoso, agressivo, cansado) poderiam ser descritos através da combinação de um conjunto de actividades simples, protagonizadas por diferentes partes do corpo, a que os autores chamaram de acções de baixo nível. Assim, os movimentos dos lábios (bocejar, falar, rir),

dos olhos (pestanejar) e o movimento global do corpo (sentar, levantar, estar presente ou ausente), constituíram as acções utilizadas pelo sistema na aferição de comportamentos. Cada uma destas acções é identificada por um classificador dedicado, de baixo nível.

A classificação de comportamentos foi implementada com recurso a *BNs*, onde cada rede representa um determinado comportamento. As topologias das *BNs* derivam exclusivamente do conhecimento que os peritos de segurança tinham acerca de eventos perigosos. As *BNs* consistem então num conjunto de nós representando as acções de baixo nível. Os nós são conectados por ligações acíclicas direccionadas, expressando quantitativamente as probabilidades condicionais dos nós e seus antecessores.

No âmbito do projecto *Context Aware Vision using Image-based Active Recognition* (*CAVIAR*) [CAVIAR, 2006], foi proposto um algoritmo [Nascimento et al., 2005] para classificação de quatro actividades humanas, a partir de sequências de imagens obtidas num ambiente de um centro comercial. Essas actividades, representadas pela trajectória do centro de massa dos objectos ao longo do tempo, consistem na "*entrada*" e "*saída*" de um estabelecimento comercial, "*passar em frente*" e "*parar para observar*" a montra de uma loja. A abordagem proposta para o reconhecimento das actividades implica uma definição prévia do ambiente monitorizado.

Jacinto Nascimento *et al.* demonstraram que, através de cinco acções elementares (i.e., "*mover para cima*", "*parar*", "*mover para baixo*", "*mover para a esquerda*" e "*mover para a direita*"), era possível representar actividades. A filosofia desta abordagem assenta na ideia de que uma actividade pode ser decomposta em acções simples, o que facilita o seu processo de classificação.

Cada acção elementar é modelada por um *cluster*, representado por um *GM* e obtido através de um conjunto de dados de treino. Cada *cluster* é composto pelos deslocamentos efectuados pelos objectos, durante uma acção elementar, obtidos de imagens sucessivas. As actividades observadas são posteriormente classificadas através de comutações verificadas entre modelos de acções elementares. Como exemplo, a actividade "*passar em frente*" seria descrita apenas por um modelo, i.e., "*mover para a esquerda*" ou "*mover para a direita*". Por outro lado, a actividade "*entrar*" poderia ser descrita por dois modelos, e.g. "*mover para a direita*" seguido por "*mover para cima*".

2.3. Discussão

A identificação e previsão de comportamentos de objectos (e.g. pessoas e veículos) observados em imagens digitalizadas e não comprimidas, abrange, como se verificou neste levantamento do trabalho relacionado, duas áreas de investigação distintas, i.e., a *Visão por Computador* e a *Inteligência Artificial*.

A visão por computador está na génese de todo o sistema. É com recurso a estas técnicas que se efectuam a aquisição, processamento e análise de sequências de imagens. Através da utilização de algoritmos de visão por computador espera-se segmentar, seguir e identificar diferentes tipos de objectos. O objectivo consiste em reunir um conjunto de dados, de diferentes atributos, que caracterize o comportamento de cada objecto durante o período de tempo em que este se mantenha em cena.

2.3.1. Segmentação de Objectos

Com a segmentação de objectos em movimento, espera-se definir as regiões de pontos de uma imagem, relevantes para descrever as componentes de forma e cor de todos os objectos não pertencentes ao plano de fundo de uma cena. Um método de segmentação ideal deveria ser suficientemente robusto contra alterações ambientais, mas sensível o bastante para detectar os objectos de interesse em movimento.

Constata-se que apesar de constituir uma área com intensa actividade científica, o problema da segmentação de objectos em movimento em ambientes complexos está longe de ser completamente resolvido. A variação de luminosidade ao longo do dia, ou devido à oclusão solar por nuvens, as condições atmosféricas adversas como o nevoeiro, a chuva, a neve e o vento, bem como as sombras dos objectos em movimento, são alguns dos inúmeros factores indutores de ruído no processo de segmentação.

Embora tenham sido propostos inúmeros métodos de segmentação de objectos em movimento, a grande maioria dos trabalhos efectuados nesta área podem ser coligidos em três grupos: **diferenciação temporal**; **fluxo óptico**; e **subtracção do plano de fundo**.

As técnicas de segmentação baseadas na **diferenciação temporal** identificam objectos em movimento através do cálculo da diferença absoluta entre imagens sucessivas, separadas por um curto intervalo de tempo. As imagens processadas podem divergir quanto ao

espaço de cor adoptado, podendo ser consideradas imagens em escala de cinzentos ou em cor (e.g. *RGB, HSV, HSL*).

Na diferenciação temporal um ponto é assinalado como pertencente a uma região de primeiro plano, i.e. segmentado, se a diferença absoluta do valor desse ponto entre duas imagens contíguas for maior que um predeterminado valor, designado por *threshold*. A escolha do valor de *threshold* é, na maioria dos casos, efectuada empiricamente, podendo ter o mesmo valor para todos os pontos da imagem. Outras variantes desta técnica, explorando a informação relativa à variação de intensidade dos pontos ao longo do tempo, adoptam métodos de cálculo de valores de *threshold* dinâmicos e específicos para cada ponto da imagem.

A principal vantagem deste tipo de segmentação prende-se com a sua simplicidade. Exigindo escassos recursos computacionais, a segmentação por diferenciação temporal entre imagens é passível de ser utilizada em sistemas com requisitos temporais muito exigentes, mas cuja precisão na extracção de objectos em movimento não seja um factor crítico.

Fruto da sua extrema simplicidade, este método apresenta importantes limitações que o tornam inaceitável em sistemas com requisitos de elevada precisão na detecção de objectos em movimento. As limitações deste tipo de segmentação devem-se à existência, nos seus resultados, de áreas consideráveis de falsos positivos e falsos negativos. Neste contexto, como falso negativo entendem-se todos os pontos erradamente classificados como não pertencendo ao objecto em movimento. Por sua vez, os falsos positivos dizem respeito aos pontos seleccionados como pertencentes a um objecto em movimento, mas que na realidade não pertencem a esse objecto.

Neste tipo de abordagem, a existência de áreas no interior dos objectos com cores uniformes provoca a ocorrência dos falsos negativos. Por outro lado, os falsos positivos são gerados pela deslocação do objecto na cena, deixando atrás de si uma área que, apesar da presença efectiva movimento, não pertence ao objecto. A presença de sombras, provocadas por objectos em movimento, origina também falsos positivos.

A necessidade de ajuste do valor de *threshold* para situações com diferentes condições de luminosidade e características cromáticas dos objectos a segmentar, e de este processo

exigir intervenção humana, desencoraja também o uso desta técnica em aplicações de vídeo-vigilância reais.

As técnicas de segmentação baseadas no cálculo do **fluxo óptico** apresentam uma abordagem bem mais complexa, analisando imagens numa sequência de modo a aferir, para cada ponto, o movimento detectado (i.e., a sua velocidade e direcção). Assim, o movimento de todos os pontos de uma imagem é representado por um campo de vectores de fluxo, onde cada vector de fluxo indica a velocidade e direcção do deslocamento do ponto respectivo.

Aquando do movimento de translação de um corpo rígido, o movimento de todos os pontos da imagem, pertencentes a esse objecto, experimentam um deslocamento com direcção e intensidade similares. Explorando esse facto, inúmeras técnicas de segmentação por fluxo óptico realizam a extracção de regiões de primeiro plano da imagem, i.e. regiões pertencentes a objectos em movimento, através da junção de pontos de uma imagem com vectores de fluxo semelhantes.

O fluxo óptico é uma técnica que apresenta como principal vantagem a segmentação de objectos em movimento, monitorizados com o auxílio de câmaras móveis. Este facto tem sido explorado fundamentalmente em sistemas de navegação autónomos como o que é apresentado em [Kruger et al., 1995; Giachetti et al., 1998]. Contudo, neste caso é necessário recorrer a métodos que permitam eliminar o ego-movimento, i.e., fluxo óptico induzido pelo movimento da câmara.

A aplicação do fluxo óptico para a segmentação de objectos em movimento, em tarefas de vídeo-vigilância, debate-se porém com algumas limitações. A necessidade de processar imagens separadas por curtos espaços de tempo, normalmente de quarenta milissegundos, entra em conflito com os elevados recursos computacionais necessários para o cálculo dos vectores de fluxo. Apesar de constituir uma limitação actual, a crescente evolução da capacidade de processamento aponta para a atenuação deste problema, podendo mesmo deixar de constituir um entrave à utilização desta técnica, num futuro próximo. Mas mesmo tendo em vista um horizonte de implementação mais longínquo, onde se possa considerar que a complexidade de computação do fluxo óptico não é relevante na escolha da técnica de segmentação, existem outros problemas associados ao cálculo do fluxo óptico. Com efeito e em particular em aplicações de vídeo-vigilância, podem ser considerados factores impeditivos a sensibilidade a alterações de luminosidade, a necessidade de brilho constante

entre imagens, a falta de robustez nos casos em que sombras são geradas por objectos em movimento e os vectores de fluxo díspares originados por objectos deformáveis em movimento.

Os métodos de segmentação baseados na **subtracção do plano de fundo** são frequentemente utilizados em aplicações de vídeo-vigilância. A sua filosofia assenta na manutenção de uma imagem de referência, denominada por "plano de fundo", calculada ou adquirida aquando do arranque do sistema. Neste tipo de segmentação, a cada nova imagem capturada, é efectuada a comparação com a imagem de referência, para que regiões que evidenciem alterações sejam classificadas como regiões de primeiro plano, efectuando-se posteriormente (em algumas abordagens) uma actualização selectiva do valor dos pontos da imagem de referência de forma a incorporar pequenas variações não identificadas como movimento.

O modo como se processa a identificação da ocorrência de movimento para cada ponto da imagem é um factor relevante e com forte impacto no sucesso da segmentação. Frequentemente, assinala-se movimento em pontos cuja diferença absoluta entre o valor da imagem actual e da imagem de referência, para esse ponto, é superior a um determinado valor, denominado de *threshold*. A escolha do valor de *threshold* implica um compromisso na sensibilidade da detecção. Este deve ser suficientemente alto para ignorar ruído, mas ao mesmo tempo baixo para permitir detectar correctamente os objectos em movimento.

No contexto desta técnica encontramos diversas variantes, cada uma apresentado as suas vantagens e limitações. Estas variantes divergem na escolha do modelo de representação do plano de fundo, do modo como se processa a identificação de movimento em cada ponto, bem como na técnica adoptada para a actualização do plano de fundo (quando aplicável).

A modelação do plano de fundo através de uma distribuição gaussiana [Buzan, 2004], implica o cálculo de duas máscaras (máscara de média e máscara de desvio padrão) que definem um valor médio e respectivo intervalo de variação, admissíveis para que cada ponto da imagem se considere pertencer ao plano de fundo. Tipicamente, este tipo de abordagem requer, aquando do arranque do sistema, um período de aprendizagem da imagem de referência, durante o qual se adquirem imagens desprovidas de objectos em movimento. Em alguns casos, a imagem de referência é adaptada ao longo do tempo, assimilando informação e estatísticas extraídas das imagens capturadas.

Baseando-se nas estatísticas da imagem de referência, as técnicas que adoptam uma modelação do plano de fundo por uma distribuição gaussiana, tomam como valor de *threshold*, para cada ponto, um múltiplo do seu desvio-padrão (normalmente duas ou três vezes o valor do desvio-padrão). Possibilita-se, deste modo, que diferentes regiões da imagem tenham a sensibilidade da detecção ajustada à usual variação da luminosidade e cor em cada ponto.

Através de uma correcta definição do período de aprendizagem do plano de fundo, é possível dotar o sistema com informação que lhe permita distinguir movimentos cíclicos originados por equipamentos pertencentes ao plano de fundo (como por exemplo escadas rolantes ou árvores abanando com o vento), de movimentos de objectos sobre essas áreas (e.g. pessoa deslocando-se numa escada rolante). Contudo, para além de em situações reais ser muitas vezes impossível ter um controlo sobre as movimentações dos objectos em cena, i.e., adquirir imagens sem objectos populando a cena, o período de aprendizagem, que por razões de natureza de implementação tem uma duração relativamente curta, impede a assimilação de fenómenos de duração superior ao período destinado à aquisição da imagem de referência (e.g. variação da luminosidade ao longo do dia, ou oclusão da luz solar por uma nuvem).

A modelação do plano de fundo apenas por uma única distribuição gaussiana não permite reter todas as variações que ocorrem sobre o plano de fundo, devido não só à existência de ruído no sensor, mas também devido a movimentos de objectos pertencentes ao plano de fundo. Para obviar este problema, alguns autores propuseram a modelação do plano de fundo por uma mistura de gaussianos (*MOG*[12]) [Stauffer & Grimson, 1999; Power & Schoonees, 2002]. Todavia, esta nova variante acarreta um aumento da complexidade e da exigência de recursos de memória do sistema, sendo que a definição do número de distribuições gaussianas é feita *a priori*, sem ter em consideração as verdadeiras necessidades.

Uma outra variante identificada na segmentação por subtracção do plano de fundo, consiste na utilização de múltiplos planos de fundo [Boult et al., 1998; 1999; 2001]. O objectivo, semelhante à modelação por *MOG*, baseia-se na determinação de vários segmentos de valores admissíveis para cada ponto. Contudo, a semelhança entre a técnica

[12] Do inglês, *Mixture of Gaussians*.

proposta por Boult e a modelação por *MOG*, verifica-se apenas no conceito subjacente. Ao contrário da *MOG*, nesta abordagem os planos de fundo são definidos por um modelo de incremento condicional, não se calculando a variância dos pontos. Ao invés, utiliza-se uma medida simples de distância do valor do ponto, em escala de cinzentos, para determinar qual dos planos de fundo se aproxima mais do valor actual.

Aquando da actualização das múltiplas imagens de referência, o *threshold* de cada ponto é também actualizado. Assim, para cada ponto, esse valor é incrementado ou decrementado consoante se detecte um falso negativo ou um falso positivo, respectivamente. Contudo, este procedimento dinâmico de actualização do *threshold* pode originar problemas de estabilidade no processo de segmentação.

Apesar das inúmeras variantes, a segmentação de objectos em movimento através da subtracção de plano de fundo, evidencia inúmeros problemas. Sombras projectadas pelos objectos, bem como alterações súbitas de luminosidade, resultam inúmeras vezes em imagens segmentadas erradamente. Com frequência, são capturados objectos durante a inicialização do sistema (e.g. automóveis parados num parque de estacionamento) que erradamente se assume pertencerem ao plano de fundo. A ausência de mobilidade por longos períodos de tempo, de objectos que outrora se encontravam em movimento, originará provavelmente a sua incorporação na imagem de referência. Devido a estas particularidades, surge um outro género de erro, associado unicamente a este tipo de abordagem: a existência de *fantasmas* nos resultados da segmentação. Estes *fantasmas* são gerados sempre que objectos pertencentes ao plano de fundo entram subitamente em movimento, originando erros que podem comprometer seriamente o resultado da segmentação.

Como se pode verificar, nenhum tipo de abordagem (diferenciação temporal, fluxo óptico ou subtracção do plano de fundo) satisfaz cabalmente os requisitos de segmentação de um sistema de vídeo-vigilância. Contudo, observa-se que as técnicas de subtracção do plano de fundo apresentam vantagens notórias em relação à diferenciação temporal e à utilização do fluxo óptico. Se se provar ser possível a associação de um método de segmentação por subtracção do plano de fundo com técnicas de detecção e remoção de *fantasmas* e sombras, e se todas estas operações cumprirem os requisitos temporais de um sistema de vídeo-vigilância, então, provavelmente, esta será a técnica de eleição para a segmentação de objectos em movimento.

2.3.2. Seguimento de Objectos

Após a segmentação dos objectos e para que seja possível caracterizar os seus comportamentos, é imperativo que se efectue o seu seguimento ao longo do tempo, adquirindo a cada instante os atributos que se julguem adequados para a representação de actividades e sua posterior padronização. Observou-se que a tarefa de seguir objectos em movimento evoluiu desde as mais rudimentares técnicas em que se associavam regiões que sobrepunham espacialmente regiões segmentadas na imagem anterior [Yamamoto et al., 1995], ou que possuíam centros-de-massa a uma distância inferior a um determinado valor [Cucchiara et al., 2001], até métodos mais complexos que foram posteriormente propostos, como os *Filtros Kalman* [Stauffer & Grimson, 2000; Buzan, 2004], *Modelos de Aparência* [Haritaoglu et al., 2000], *Redes Bayesianas* [Jorge et al., 2004], *Hidden Markov Models* e, mais recentemente, os *Filtros de Partículas* [Chang & Ansari, 2005].

A oclusão parcial ou total dos objectos em movimento, a junção de diferentes objectos bem como separação de objectos de um grupo, representam os principais desafios que se levantam ao processo de seguimento. As tarefas de vídeo-vigilância implicam não só o seguimento de objectos rígidos (e.g. automóveis), mas também objectos deformáveis (e.g. pessoas e animais). Em particular, no caso do seguimento de pessoas, a sua forma e cor podem-se alterar drasticamente ao longo do seu percurso pela cena.

Verifica-se que as abordagens mais elementares, que executam o seguimento de objectos através da sobreposição de áreas ou pela distância do centro-de-massa de um objecto em imagens sucessivas, não asseguram resultados robustos. Estas técnicas falham nos casos em que dois ou mais objectos se cruzam ou formam um grupo. Nestas situações, torna-se impossível distinguir individualmente os objectos.

A utilização de *Filtros Kalman* no seguimento de objectos evidencia limitações devidas ao facto de se basear em densidades gaussianas unimodais, técnica esta que não suporta hipóteses alternativas simultâneas de movimento [Isard & Blake, 1996]. A utilização de *Filtros Kalman Estendidos* contorna este problema, através da manutenção de uma lista de múltiplas hipóteses nos casos onde se verifique uma ambiguidade entre múltiplos objectos em movimento. Todavia, este tipo de solução continua a não proporcionar resultados suficientemente robustos em situações reais de maior complexidade.

O recurso a *Modelos de Aparência* [Haritaoglu et al., 2000] para o seguimento de objectos ao longo do tempo demonstrou um elevado grau de fiabilidade e robustez quando se trata de situações de oclusão total ou parcial. Este tipo de técnica garante sobretudo resultados superiores quando o modelo dos objectos é composto pela combinação de dois tipos de informação: uma *Imagem de Aparência* e uma *Máscara de Probabilidade* [Senior, 2002; Cucchiara et al., 2004]. Os *Modelos de Aparência* podem ser vistos como representações dinâmicas dos objectos, que são mantidas em memória pelo sistema, e actualizadas ao longo do tempo, de modo a assimilar as alterações sofridas na sua aparência e forma ao longo do percurso. Assim, a *Imagem de Aparência* mantém um modelo de cor do objecto, enquanto que a *Máscara de Probabilidade* apresenta a forma mais provável que o objecto pode assumir.

Os *Modelos de Aparência*, graças ao seu carácter dinâmico, apresentam-se como uma solução adequada para o seguimento de objectos em tarefas de vídeo-vigilância. Sobretudo porque esta técnica tira partido das informações de mutação de forma e cor, próprias da locomoção dos seres humanos. Apesar das inquestionáveis vantagens, a utilização deste tipo de técnica requer elevados recursos computacionais e de memória, uma vez que necessita de manter e actualizar os modelos de todos os objectos observados numa cena.

A utilização de *Redes Bayesianas* para o seguimento de objectos em movimento obteve resultados satisfatórios, como os apresentados por [Jorge et al., 2004]. Através de um processo de redução gradual da influência do histórico de informações na tomada de decisões, evitou-se a explosão de combinações, mantendo a complexidade da rede dentro de limites razoáveis, o que possibilitou a utilização tempo-real deste método sobre sequências de imagens de longa duração. Tal não acontecia com os anteriores modelos baseados neste tipo de redes [Abrantes et al., 2002].

Apesar dos avanços obtidos na utilização de *Redes Bayesianas*, este tipo de abordagem carece de universalidade na sua aplicação. Esta limitação deve-se ao facto de a construção de uma rede implicar a definição do número fixo de nós que a compõem e, devido a esta restrição, apenas se poder manter informação acerca de um período de tempo, mais ou menos limitado, consoante o número de nós permitido. Assim, a correcta resolução de conflitos originados por fenómenos de oclusão, junção ou separação de objectos, dependerá sempre da informação adquirida e mantida pela rede. Erros no seguimento de objectos, outrora resolvidos com sucesso, poderão ocorrer nos casos em que a informação tenha sido eliminada da rede.

Analisando as diferentes abordagens propostas para o seguimento de objectos em movimento, observa-se que a utilização de *Modelos de Aparência* apresenta vantagens notórias em relação às restantes soluções apresentadas. A sua filosofia é bastante próxima do que acontece com o reconhecimento de objectos pelos seres humanos, uma vez que é mantido em memória o modelo de cada objecto. A analogia com o processo humano de reconhecimento de objectos pode ser feita, por exemplo, através da situação em que se exibe uma fotografia de um determinado objecto a um indivíduo, pedindo-lhe de seguida para o identificar de entre um conjunto de objectos. O indivíduo seria capaz de identificar o referido objecto, mesmo que este se encontrasse com uma posição ou forma diferente da que lhe fora anteriormente exibida, pois simplesmente seleccionaria o objecto que se lhe afigurasse mais provável.

Contudo, apesar das vantagens apresentadas não se espera que os *Modelos de Aparência* constituam uma panaceia para todos os problemas próprios do seguimento de objectos em situações reais. Nas aplicações de seguimento de atletas, em desportos colectivos, como por exemplo o futebol, todos os jogadores de uma equipa (à excepção do guarda-redes) apresentam uma forma e cor semelhantes. Neste tipo de situação as técnicas baseadas em *Modelos de Aparência* não assegurariam resultados satisfatórios aquando da oclusão de um atleta por um elemento da mesma equipa. Este problema aconteceria ainda que os atletas se deslocassem em sentidos opostos.

2.3.3. Análise de Movimento

Recorrendo a técnicas de inteligência artificial, várias abordagens foram propostas para a identificação e previsão de actividades humanas. Modelos não estocásticos, modelos discriminativos, modelos estocásticos descritivos e modelos estocásticos generativos, têm vindo a ser aplicados na identificação de padrões de comportamento. Apesar da ampla diversidade de técnicas propostas até ao momento, nenhuma das abordagens mostrou resultados suficientemente robustos e precisos na realização dessa tarefa.

Os modelos não estocásticos impossibilitam que se efectue uma aprendizagem automática dos padrões de comportamento. Esta característica implica que os modelos de comportamento sejam definidos pelo utilizador do sistema, recaindo sobre este a tarefa de identificar e modelar as actividades que se pretendam observar. Entre as técnicas mais populares de modelação não estocástica de comportamentos encontram-se as *Finite State*

Machines (*FSMs*) [Gill, 1962], a *Dynamic Time Warping* (*DTW*) [Sakoe & Chiba, 1978; Rabiner et al., 1978], e a *Longest Common Subsequence* (*LCSS*) [Cormen et al., 1990].

Os modelos discriminativos possibilitam já a realização da aprendizagem automática de padrões de comportamento. Para tal, estas técnicas necessitam de um conjunto de dados de treino, devidamente classificados segundo as actividades a identificar, de modo a mapear, de uma forma directa, um vector de entrada em limites de superfícies de decisão. Devido à frequente sobreposição de classes em determinadas regiões do domínio de soluções, as abordagens baseadas em modelos discriminativos recorrem a funções discriminantes, de modo a definir uma medida de separação entre classes.

De entre os modelos discriminativos mais relevantes encontram-se as *Neural Networks* (*NNs*) [Davalo, 1988], o *K-Nearest Neighbour* (*K-NN*) [Cover & Hart, 1967], o *Self-Organizing Map* (*SOM*) [Kohonen, 1997], o *Support Vector Machines* (*SVMs*) [Burges, 1998], o *Maximum Entropy Markov Model* (*MEMM*) [McCallum et al., 2000], e o *Conditional Random Fields* (*CRF*) [Lafferty et al., 2001].

Os modelos estocásticos descritivos permitem também efectuar a aprendizagem de comportamentos. Contudo, este tipo de abordagem não necessita de supervisão durante o processo de aprendizagem. Este facto deve-se à filosofia deste tipo de modelação que pretende descobrir relações entre os dados, descrevendo as principais características, como os modelos de distribuição de probabilidades, o seccionamento de um hiperespaço em diversos grupos, ou pela modelação de dependências entre atributos.

Técnicas como o *K-Means* [MacQueen, 1967], o *Expectation-Maximization* (*EM*) [Dempster et al., 1977], o *Numeric Iterative Hierarchical Cluster* (*NIHC*) [Wallace, 1989], e o *O-Cluster* [Milenova & Campos, 2002], são exemplos de implementação de modelos estocásticos descritivos.

As técnicas de classificação e previsão de comportamentos baseadas em modelos estocásticos generativos permitem gerar modelos estatísticos de comportamentos futuros. O objectivo consiste na definição de um modelo específico que interprete o modo como as observações são geradas. Deste modo, assume-se a existência de um conjunto de variáveis ocultas, cuja organização se desconhece, que são responsáveis pela geração dos dados observados.

A identificação das variáveis ocultas e do modo como estas se relacionam, tem vindo a ser explorado, em diversas técnicas, tais como a *Principal Components Analysis* (*PCA*) [Jolliffe, 1986], os *Modelos de Misturas de Gaussianos* (*GMM*) [McLachlan & Basford, 1988], os *Hidden Markov Models* (*HMM*) [Rabiner, 1989], e as *Redes Bayesianas* (*BN*) [Duda et al., 2001].

As especificidades dos sistemas de vídeo-vigilância implicam que as abordagens a utilizar na classificação e previsão de comportamentos permitam operar por longos períodos de tempo, idealmente de forma ininterrupta, e sobre uma vasta variedade de condições de operação. Em particular, estas técnicas deverão estar dotadas de mecanismos que lhes permitam adquirir e processar um elevado número de observações em tempo-real.

Adicionalmente, as técnicas de aprendizagem devem ser capazes de operar automaticamente, i.e., sem intervenção humana, e permitir a adaptação dos modelos de comportamento, reflectindo de forma instantânea os comportamentos observados nos dados que vão sendo adquiridos ao longo do tempo. Deste modo, todas as técnicas de aprendizagem supervisionada, que requeiram uma classificação manual de conjuntos de dados de treino e cuja adaptação a novas condicionantes exija intervenção humana, não cumprem os requisitos dos sistemas de vídeo-vigilância. Assim, apenas métodos baseados em aprendizagem não supervisionada serão apropriados nas tarefas de aprendizagem de comportamentos em tarefas de vídeo-vigilância.

Capítulo 3

3. Modelos de Representação de Cor

Os modelos de representação de cor permitem caracterizar de uma forma quantitativa qualquer cor definida pelo espectro de luz visível, i.e. com comprimento de onda entre 400nm e 700nm. Para tal, cada modelo ou espaço de cor está associado a uma função que permite mapear a **distribuição da potência espectral** (*DPE*) da radiação electromagnética visível, num espaço definido por um conjunto de valores discretos que quantificam as componentes de cor que compõem o modelo.

Certos espaços de cor mostram-se sensíveis a variações das condições de iluminação, enquanto que outros garantem a preservação de determinadas características cromáticas, permanecendo imunes a essas alterações. Assim, torna-se necessário identificar as vantagens e fraquezas de cada modelo, por forma a fundamentar a adopção dos espaços de cor a utilizar nas técnicas de processamento e análise de imagem no âmbito deste trabalho.

Este capítulo abre com algumas considerações sobre o vídeo e a imagem digital, apresentando as principais normas e formatos. Prossegue-se com a definição da resposta do sensor de aquisição de imagem, descrita pelo modelo de reflexão dicromático. Em seguida, os espaços de cor *rgb*, $c_1c_2c_3$, $l_1l_2l_3$ e *HSV* são avaliados tendo como objectivo a aferição da sua invariabilidade a sombras e brilhos. O presente capítulo encerra com uma discussão sobre as potencialidades dos espaços de cor analisados.

3.1. Vídeo Digital

O vídeo digital é composto pela sequência contínua de imagens digitais estáticas, habitualmente designadas por fotogramas (como representado na Figura 3.1), adquiridas em intervalos de tempo fixos e suficientemente curtos, de modo a criar nos observadores humanos a percepção de movimento ininterrupto. Numa câmara de vídeo, analógica ou digital, as imagens são adquiridas por um sensor *CCD*[13] ou *CMOS*[14] que gera, para cada amostragem, uma representação analógica da imagem que será posteriormente convertida para a forma digital (na própria câmara no caso de câmaras digitais, ou numa placa de aquisição de vídeo, quando se empregam câmaras analógicas).

Figura 3.1. Exemplo de uma sequência de imagens digitalizadas que compõem um vídeo.

Nos sistemas híbridos, em que a arquitectura se baseia na utilização de câmaras de vídeo analógicas, cujo sinal é transmitido a uma placa de aquisição de vídeo acoplada a uma unidade de processamento, a resolução da imagem e as suas características cromáticas dependem do formato adoptado pelo equipamento.

Existem actualmente diversos formatos de aquisição de vídeo, sendo os mais utilizados em sistemas de vídeo-vigilância o *PAL* [Jack, 2005] e o *NTSC* [NTSC, 1954; Jack, 2005]. Cada formato especifica uma diferente velocidade de captura de imagens que é medida em unidades de "imagens/segundo" ou em inglês *frames per second* (*f.p.s.*), define a divisão da imagem num determinado número de linhas, indica uma resolução máxima permitida, e descreve um sistema de cor (*YUV* para o formato *PAL*, e *YIQ* para o *NTSC*).

Apesar das características distintivas de cada formato, estas não apresentam qualquer relevância na escolha e implementação das técnicas de processamento e análise de imagem utilizadas para a segmentação e seguimento de objectos em movimento. Independentemente do formato utilizado, cada imagem adquirida por uma câmara de vídeo

[13] Do inglês, *Charged Couple Device*.

[14] Do inglês, *Complementary Metal Oxide Semiconductor*.

é digitalizada, pela placa de captura, de acordo com um espaço de cor preestabelecido, normalmente o espaço de cor *RGB*. A transformação para este espaço, a partir dos espaços *YUV* e *YIQ* é obtida pelas seguintes operações [Andleigh & Thakrar, 1996]:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.140 \\ 1 & -0.395 & -0.581 \\ 1 & 2.032 & 0 \end{bmatrix} \cdot \begin{bmatrix} Y \\ U \\ V \end{bmatrix} \tag{3.1}$$

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & 0.956 & 0.621 \\ 1 & -0.272 & -0.647 \\ 1 & -1.105 & 1.702 \end{bmatrix} \cdot \begin{bmatrix} Y \\ I \\ Q \end{bmatrix} \tag{3.2}$$

Com a utilização de câmaras digitais, o sinal de vídeo é convertido para o domínio digital nas próprias câmaras, representado frequentemente em maior detalhe pelo espaço de cor *RGB*. As câmaras digitais possuem já consideráveis capacidades de processamento que proporcionam diversas vantagens. Pelo facto da conversão do sinal analógico para o formato digital ser realizada próxima do sensor, o ruído electrónico induzido no sinal de vídeo é reduzido ao mínimo. Um outro benefício resulta de que, após a digitalização, o sinal de vídeo transmitido (e.g. sobre rede *ethernet*) se torna imune a ruído, permitindo assim que a câmara esteja afastada do equipamento receptor de imagens, sem que esse facto implique qualquer degradação do sinal.

As câmaras digitais apresentam ainda uma outra vantagem adicional em relação às arquitecturas híbridas, uma vez que não manifestam flutuações na imagem causadas durante recuperação do sincronismo, problema também conhecido por *pixel jitter*, que ocorre frequentemente nas placas de aquisição de vídeo. Este tipo de flutuações, mesmo que desprezável ao olho humano, é causador de dificuldades no processamento e análise das imagens.

Em suma, verifica-se que independentemente da arquitectura adoptada pelo sistema de vídeo-vigilância (híbrido ou digital), é sempre possível obter uma sequência de imagens digitalizadas, que caracterizem com o máximo de informação a cena observada. A utilização do espaço de cor *RGB* para a representação das imagens, permite definir uma base comum, garantindo a compatibilidade entre qualquer tipo de tecnologia adoptada para a aquisição de vídeo e as técnicas de processamento de imagem situadas a jusante, que venham a ser desenvolvidas.

3.1.1. Imagem Digital

Em processamento digital de imagem, quer se trate da análise de apenas uma única fotografia ou de uma sequência de vídeo constituída por um elevado número de fotogramas, o alvo de processamento é a imagem e os dados nela contidos. Para que se possa efectuar, de uma forma coerente, as operações de processamento e análise sobre as imagens, torna-se necessário definir uma notação para a descrição das imagens, bem como da sua estrutura.

Neste trabalho define-se I^t como a imagem digital obtida no instante de tempo t. Uma imagem digital é uma representação em duas dimensões de uma imagem num conjunto finito de elementos que tomam valores discretos, organizada numa matriz de M por N elementos. Estes elementos, que armazenam o valor da intensidade luminosa da imagem naquela coordenada, são denominados por pontos (embora na literatura inglesa sejam definidos por *pixel*, acrónimo derivado de "*picture element*"). Assim, a intensidade de um ponto nas coordenadas x e y da imagem I^t é definida por $I^t(\mathbf{x})$, em que $\mathbf{x}=(x, y)$, $0 \leq x \leq M$, $0 \leq y \leq N$ e $I^t(\mathbf{x})$ toma valores do intervalo [0, 255].

No caso de uma imagem colorida, esta é definida pelo conjunto das várias componentes de cor. Assim, como exemplo, uma imagem do espaço de cor *RGB* é definida por $I^t=\{I_R^t, I_G^t, I_B^t\}$, em que $I_R^t(\mathbf{x})$, $I_G^t(\mathbf{x})$ e $I_B^t(\mathbf{x})$ representam respectivamente o valor da intensidade das componentes vermelha, verde e azul, para o ponto definido pelas coordenadas (x, y), da imagem digital adquirida no instante de tempo t.

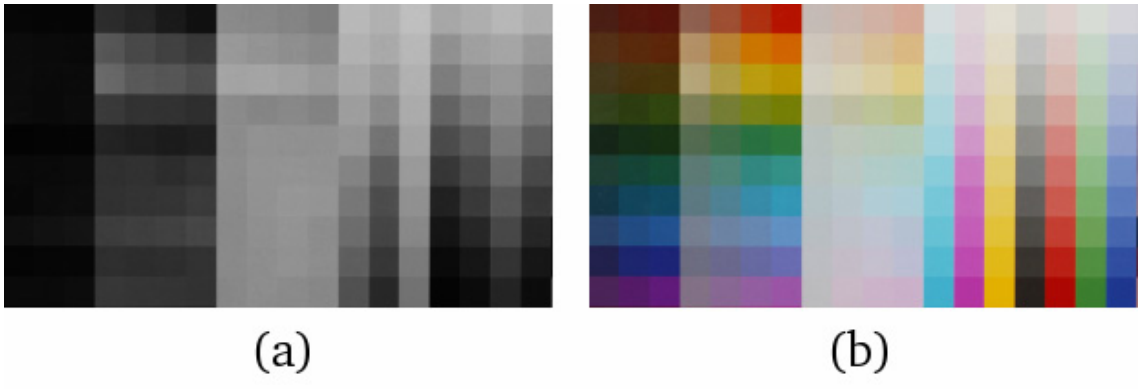


Figura 3.2. (a) Imagem em escala de cinzentos; (b) Imagem colorida, definida pelo espaço de cor *RGB*.

3.1.2. Espaço de Cor *RGB*

Um modelo ou espaço de cor proporciona um método formal de representação das cores. Através da adopção de um sistema de coordenadas tridimensional, que represente valores de três componentes ortogonais de um espaço de cor, é possível representar qualquer cor do espectro visível, possibilitando igualmente o emprego de técnicas analíticas sobre esse espaço de cor. Um modelo simples de representação de cor, baseado na forma de como o olho humano a adquire [Dowling, 2002], é o espaço de cor *RGB*.

O espaço de cor *RGB* consiste num modelo de cor aditivo, onde através da combinação de três componentes primárias, i.e. vermelho (*R*), verde (*G*) e azul (*B*), é possível reproduzir qualquer outra cor do espectro de luz visível, como exemplificado na Figura 3.3. Este espaço de cor pode ser representado por um cubo, definido pelos eixos *R*, *G* e *B*, onde cada uma das componentes de cor toma valores discretos, compreendidos por exemplo entre 0 e 255 numa configuração de 24 *bits* por ponto. Na Figura 3.4 apresentam-se diferentes cores, exibindo a sua decomposição nas componentes do espaço de cor *RGB*.

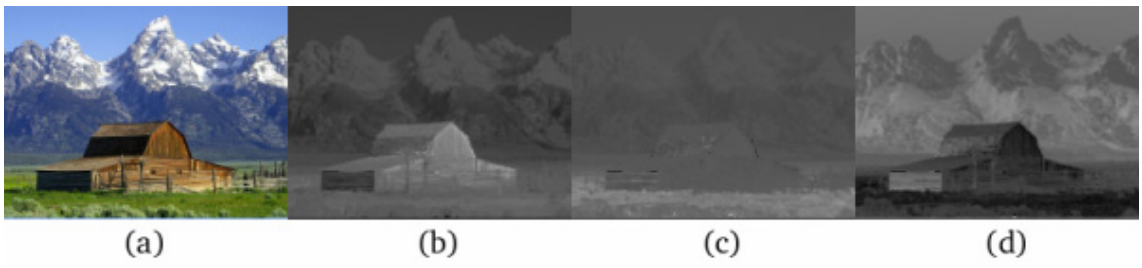


Figura 3.3. (a) Imagem em formato *RGB* e as suas componentes (b) vermelho, (c) verde, e (d) azul.

Neste modelo de cor, definido pelo cubo *RGB*, a origem representa o preto ($I_R(\mathbf{x})=0$, $I_G(\mathbf{x})=0$, $I_B(\mathbf{x})=0$), enquanto que o vértice oposto do cubo descreve o branco ($I_R(\mathbf{x})=255$, $I_G(\mathbf{x})=255$, $I_B(\mathbf{x})=255$). O eixo diagonal que percorre a origem até ao branco constitui uma zona acromática, que define uma escala de cinzentos. As magnitudes das três componentes nessa diagonal tomam o mesmo valor, i.e., $I_R(\mathbf{x})=I_G(\mathbf{x})=I_B(\mathbf{x})$. Na Figura 3.4 apresenta-se o cubo *RGB* cujo volume compreende todas as cores possíveis.

Para além da simplicidade geométrica inerente, o espaço de cor *RGB* possui ainda o benefício de requerer pouco processamento para a sua obtenção. Isto porque muitos sensores das câmaras de vídeo adquirem e apresentam a imagem nesse modelo cor.

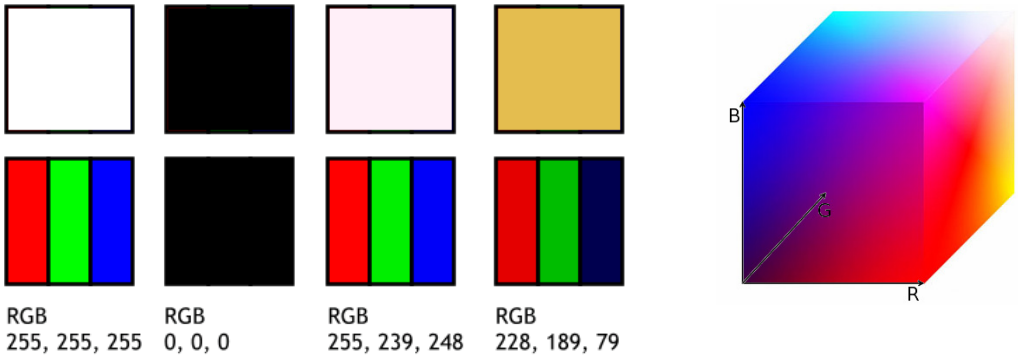


Figura 3.4. Diferentes cores (em cima), definidas pelos respectivos valores das componentes *R*, *G* e *B* (em baixo) do cubo *RGB* (à direita).

Apesar dos benefícios intrínsecos, o espaço de cor *RGB* evidencia algumas limitações. Designadamente, o modelo não exibe um comportamento uniforme em termos da percepção da cor, uma vez que uma variação de valor unitário no cubo *RGB* não se reflecte de igual modo nas três componentes do modelo.

As fragilidades deste modelo atingem maior relevância quando aplicado, por exemplo, em técnicas de segmentação de objectos, onde há a necessidade de comparar a similaridade de cores entre pontos de uma imagem e um objecto de referência. Em ambientes reais ocorrem frequentemente variações das condições de iluminação. Verifica-se que a alteração da intensidade luminosa incidente sobre a superfície de um objecto, provoca uma translação no espaço *RGB*, das coordenadas que definem a cor, que se manifesta de forma díspar nas três componentes do modelo. Assim, é comum não obter uma correspondência entre pontos de uma mesma superfície quando se alteram as condições de iluminação.

Segundo Von Kries [1902], as respostas tricromáticas medidas numa superfície sob condições de iluminação distintas (como apresentado na Figura 3.5), estão distanciadas por simples factores de escala. Assim, no espaço de cor *RGB*, considerando que ($I_R(\mathbf{x})$, $I_G(\mathbf{x})$, $I_B(\mathbf{x})$) e ($I'_R(\mathbf{x})$, $I'_G(\mathbf{x})$, $I'_B(\mathbf{x})$) representam as respostas de cor de um ponto de uma superfície sob duas condições de iluminação então, segundo a regra de Von Kries, tem-se:

$$\begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{bmatrix} \cdot \begin{bmatrix} I'_R(\mathbf{x}) \\ I'_G(\mathbf{x}) \\ I'_B(\mathbf{x}) \end{bmatrix} \tag{3.3}$$

Figura 3.5. Padrão de cores em três diferentes condições de iluminação.

Esta definição é particularmente útil nos casos em que a variação da iluminação não é proporcional nas três componentes primárias. Tais situações poder-se-iam dever, por exemplo, à aplicação de um filtro óptico que provocaria o aumento ou diminuição de um determinado componente de cor. Contudo, em ambientes exteriores tais factos não se verificam com frequência, podendo-se considerar que uma fonte de luz irradia proporcionalmente as três componentes de cor. Deste modo, a regra de Von Kries pode ser simplificada pela aplicação de um factor de escala (φ), comum a todos os componentes, relativo à variação da intensidade luminosa, onde:

$$\begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} = \varphi \cdot \begin{bmatrix} I'_R(\mathbf{x}) \\ I'_G(\mathbf{x}) \\ I'_B(\mathbf{x}) \end{bmatrix} \tag{3.4}$$

Assim, tendo por base a regra de Von Kries, um modelo invariante de representação de cor deverá possibilitar a conservação da informação da cor independentemente do nível de iluminação aplicado a uma superfície, i.e. deverá eliminar a componente de escala relacionado com a alteração das condições de iluminação (φ). A este processo dá-se o nome de cancelamento do factor φ.

A proposta de Von Kries é no entanto simplista. Numa imagem adquirida através de uma câmara de vídeo, a cor definida em cada ponto é função de diversos factores. Nomeadamente, das características do sensor, da composição da luz incidente sobre o objecto, e das propriedades reflectoras da sua superfície. Assim, surge a necessidade de definir um modelo mais abrangente, que abarque todas as condicionantes que influenciem a percepção da cor.

3.2. Modelo de Reflexão Dicromático

O modelo de reflexão dicromático foi introduzido em [Shafer, 1985] como forma de caracterizar a luz reflectida em objectos não homogéneos, como polímeros, tecidos ou rochas, que exibem dois tipos de reflexões distintas: difusa e especular. A reflexão especular consiste na componente da radiação incidente sobre a superfície de um objecto que é reflectida num determinado ângulo. De acordo com a lei da reflexão, o ângulo de reflexão terá um valor igual ao ângulo que o raio incidente faz com a normal ao ponto da superfície do objecto onde a radiação incide. Este tipo de reflexão varia de acordo com o índice de refracção do material que constitui o objecto, uma vez que, alguma radiação poderá ser absorvida pelo objecto. Por outro lado, a reflexão difusa é função da absorção da radiação pelo objecto, definindo a componente da radiação que penetra na amostra, sofrendo dispersão antes de voltar à superfície, sendo posteriormente reflectida em várias direcções.

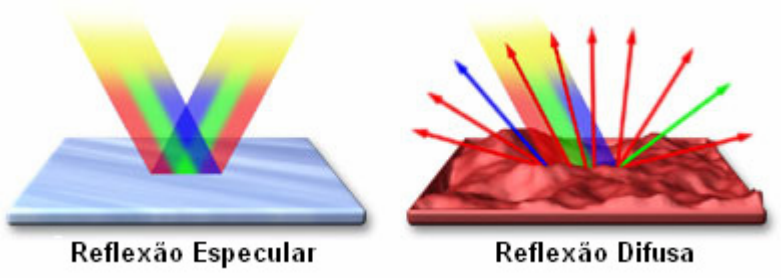


Figura 3.6. Exemplos de reflexão especular e de reflexão difusa.

A **distribuição da potência espectral** (*DPE*) da reflexão especular é idêntica à da luz que ilumina o objecto, enquanto que na reflexão difusa representa a cor que o caracteriza. A Figura 3.6 ilustra os dois tipos de reflexão.

Em superfícies reflectoras perfeitas apenas se verifica a existência de reflexão especular, enquanto que uma superfície difusa ideal, também conhecida por superfície de Lambert, se comporta como um difusor de luz perfeito. Contudo, na natureza os materiais exibem as duas componentes reflectoras, comportando maior ou menor grau de difusão ou reflexão especular. Na prática, a reflexão especular não é perfeita, sendo que a luz reflectida pode ser vista pelo observador em regiões próximas da direcção do raio de luz reflectido. Este fenómeno gera a formação de brilhos na superfície do objecto cuja intensidade varia de acordo com a posição do observador.

Considerando os dois tipos de reflexão, o modelo de reflexão dicromático descreve a luz reflectida num objecto não homogéneo como uma combinação linear das componentes de reflexão especular e difusa. Por conseguinte, a resposta do sensor à radiação emitida num ponto da superfície de um objecto não homogéneo, cuja imagem seja adquirida através de uma câmara de vídeo digital, e sobre o qual incida uma iluminação uniforme, pode ser descrita por [Lukac & Plataniotis, 2007]:

$$\mathbf{I}(\mathbf{x}) = w_d(\vec{\mathbf{n}},\vec{\mathbf{s}}) \cdot \int_{\Omega} c_d(\lambda) \cdot e(\lambda) \cdot \mathbf{q}(\lambda) \cdot d\lambda + w_s(\vec{\mathbf{n}},\vec{\mathbf{s}},\vec{\mathbf{v}}) \cdot \int_{\Omega} c_s(\lambda) \cdot e(\lambda) \cdot \mathbf{q}(\lambda) \cdot d\lambda \quad (3.5)$$

onde, $\mathbf{I}(\mathbf{x})=\{I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})\}$ define o vector de cor de um ponto na imagem, em formato *RGB*, na coordenada $\mathbf{x}=(x, y)$, para o espectro de luz visível (Ω=[400nm, 700nm]). A *DPE* da luz que incide sobre o objecto é expressa por $e(\lambda)$, sendo independente da localização espacial uma vez que se assume a existência de uma iluminação uniforme. As variáveis $w_d(\vec{\mathbf{n}},\vec{\mathbf{s}})$ e $w_s(\vec{\mathbf{n}},\vec{\mathbf{s}},\vec{\mathbf{v}})$ são factores de peso para as reflexões difusa e espectral. Ambos os factores são função da normal à superfície do objecto $(\vec{\mathbf{n}})$, e da direcção da iluminação $(\vec{\mathbf{s}})$. O factor de peso da reflexão espectral depende ainda da direcção do observador $(\vec{\mathbf{v}})$.

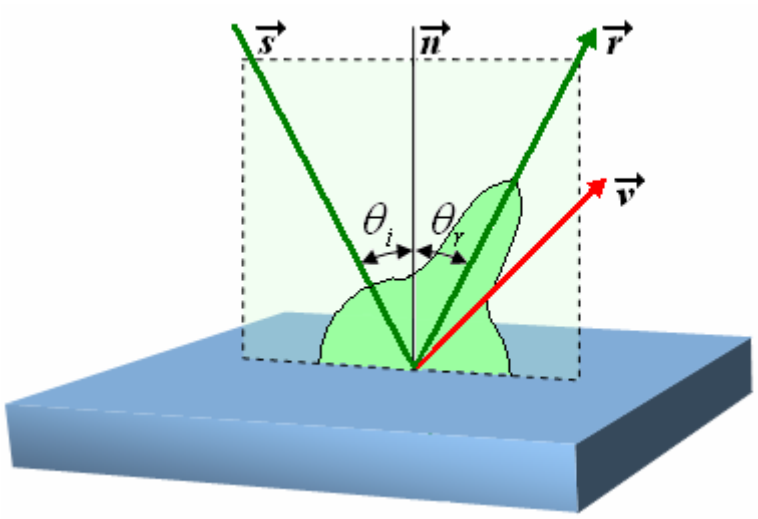


Figura 3.7. Geometria da reflexão.

Na Figura 3.7 o ângulo de incidência da radiação numa superfície está definido por θ_i, enquanto θ_r especifica o ângulo de reflexão, sendo que $\theta_i=\theta_r$. A variação da intensidade da radiação emitida para cada ângulo do plano de incidência é definida por uma parte semicircular que diz respeito à reflexão difusa, e por uma componente especular que forma um pico na região próxima do vector da reflexão especular $(\vec{\mathbf{r}})$.

O coeficiente de reflexão difusa, ou albedo[15], é representado por $c_d(\lambda)$, enquanto que o coeficiente de reflexão especular se encontra definido por $c_s(\lambda)$. A sensibilidade do sensor às componentes de cor vermelha, verde e azul são definidas pelo vector $\mathbf{q}(\lambda)=(q_R(\lambda), q_G(\lambda), q_B(\lambda))$.

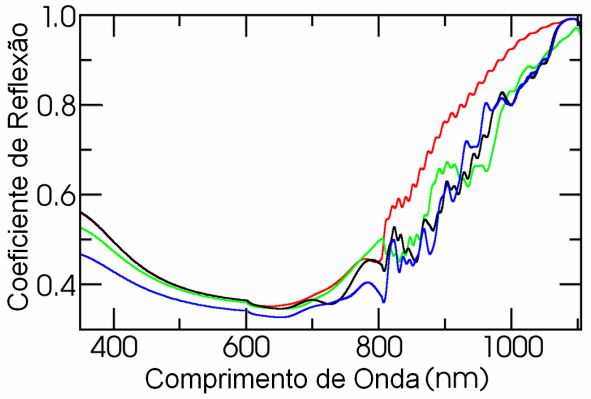


Figura 3.8. Coeficientes de reflexão difusa de quatro materiais distintos.

Desprezando o ganho e ruído da câmara de vídeo, assumindo que a luz reflectida na superfície do objecto tem aproximadamente a mesma *DPE* que a fonte de luz, sendo que essa fonte irradia igual densidade de energia em todos os comprimentos de onda do espectro de luz visível (luz branca), então, tem-se que: $e(\lambda)=E$ e $c_s(\lambda)=C_s$, i.e. são constantes no espectro de luz visível; e que:

$$\int_{\Omega} q_R(\lambda) = \int_{\Omega} q_G(\lambda) = \int_{\Omega} q_B(\lambda) = q \tag{3.6}$$

Adoptando as referidas condicionantes, a equação que define o valor medido pelo sensor da câmara pode ser reescrita na seguinte forma:

$$\mathbf{I}(\mathbf{x}) = w_d(\vec{\mathbf{n}},\vec{\mathbf{s}}) \cdot E \cdot \int_{\Omega} c_d(\lambda) \cdot \mathbf{q}(\lambda) \cdot d\lambda + w_s(\vec{\mathbf{n}},\vec{\mathbf{s}},\vec{\mathbf{v}}) \cdot C_s \cdot E \cdot q \tag{3.7}$$

Este modelo de reflexão dicromático constitui uma ferramenta útil para descrever a luz reflectida num ponto da superfície de um objecto como resultado da fusão de duas reflexões distintas. Através deste modelo é possível analisar os espaços de cor no que diz respeito às características invariantes a diversas condições, e.g. direcção do ângulo de visão e geometria do objecto, variação da intensidade da iluminação e presença de brilhos.

[15] Albedo – razão entre a energia luminosa que uma superfície difunde em todas as direcções e a luz que incide nesse elemento.

3.3. Espaços de Cor Invariantes

Em processamento digital de imagem existe muitas vezes a necessidade de recorrer a características invariantes de modo a permitir, por exemplo, identificar numa determinada imagem um objecto a partir de uma imagem modelo, ou descobrir regiões de pontos divergentes do padrão especificado por uma imagem de referência (detecção de movimento). Tais processos obrigam, entre outras operações, à computação de uma medida de similaridade, ponto a ponto, no espaço de cor estabelecido.

Devido à incapacidade de controlar o meio observado, é frequente experimentar diferentes condições de iluminação, bem como variações do ângulo de visão. Tais variabilidades produzem alterações dramáticas, se utilizado um espaço de cor sensível a variações como o modelo *RGB*, mesmo que os objectos alvo de interesse permaneçam imóveis e preservem a sua morfologia.

A título de exemplo, considere-se o seguinte cenário, em que se pretende detectar movimento numa sequência de imagens adquiridas num ambiente exterior, utilizando o espaço de cor *RGB*. Para o efeito, procede-se numa primeira fase à aquisição de uma imagem de referência, na qual apenas devem constar objectos relativos ao plano de fundo. Após a obtenção dessa imagem de referência, prossegue-se com o processamento da sequência de vídeo. Assim, para cada imagem da sequência, é então calculada, para cada ponto, uma medida de similaridade com o correspondente ponto da imagem de referência. Se durante a aquisição da sequência se verificasse uma oscilação das condições de iluminação, por exemplo através da obstrução da luz solar pela passagem de uma nuvem, possivelmente seriam detectadas consideráveis regiões de movimento, isto mesmo que todos os objectos permanecessem imóveis.

Como resposta a estas limitações, torna-se necessário recorrer a modelos de cor, invariantes às condições de iluminação (sombras e brilhos), que possuam ainda características de invariabilidade em relação ao ângulo de visão e geometria dos objectos. Assim, um modelo de cor invariante deve permitir reduzir ao mínimo absoluto a informação irrelevante, enquanto retém a maior quantidade possível de características discriminatórias.

3.3.1. Invariabilidade à Intensidade da Iluminação

É possível aferir analiticamente da invariabilidade de um espaço de cor em relação à intensidade da iluminação (na qual se inclui as sombras), através da aplicação do modelo de reflexão dicromático. Para tal, considera-se a resposta do sensor à luz reflectida num ponto de uma superfície de Lambert iluminada por uma fonte de luz branca (fonte de luz que irradia igual densidade de energia em todos os comprimentos de onda do espectro de luz visível).

De acordo com a equação (3.7), e tendo em consideração as condições referidas, a reflexão de materiais não homogéneos sobre luz branca é descrita por [Lukac & Plataniotis, 2007]:

$$\begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} = w_d(\vec{\mathbf{n}},\vec{\mathbf{s}}) \cdot E \cdot \int_{\Omega} c_d(\lambda) \cdot \begin{bmatrix} q_R(\lambda) \\ q_G(\lambda) \\ q_B(\lambda) \end{bmatrix} \cdot d\lambda \tag{3.8}$$

De modo a obter uma formulação mais compacta, considera-se que:

$$\begin{bmatrix} K_R \\ K_G \\ K_B \end{bmatrix} = \int_{\Omega} c_d(\lambda) \cdot \begin{bmatrix} q_R(\lambda) \\ q_G(\lambda) \\ q_B(\lambda) \end{bmatrix} \cdot d\lambda \tag{3.9}$$

Assim, a equação (3.8) pode ser reescrita por:

$$\begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} = w_d(\vec{\mathbf{n}},\vec{\mathbf{s}}) \cdot E \cdot \begin{bmatrix} K_R \\ K_G \\ K_B \end{bmatrix} \tag{3.10}$$

A resposta do sensor é então condicionada pelo factor geométrico da reflexão difusa, pela *DPE* da luz incidente sobre o objecto, e pela sensibilidade do sensor ao espectro de luz que caracteriza esse objecto ($K_{R,G,B}$). Um espaço de cor invariante a alterações da intensidade luminosa, apenas deve ser função dos elementos K_R, K_G e K_B, cancelando os restantes factores.

3.3.2. Invariabilidade a Brilhos

A resposta do sensor à iluminação reflectida num ponto de um objecto, sobre a qual incide uma luz branca, e cuja superfície apresenta uma forte componente especular, encontra-se plenamente definida pela equação (3.7).

De modo a obter uma formulação mais compacta, considera-se que:

$$S = w_s\left(\vec{\mathbf{n}},\vec{\mathbf{s}},\vec{\mathbf{v}}\right)\cdot C_s \cdot q \tag{3.11}$$

Através da aplicação das expressões definidas em (3.9) e (3.11), a equação (3.7) pode ser reescrita por:

$$\begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} = E \cdot \left(w_d\left(\vec{\mathbf{n}},\vec{\mathbf{s}}\right) \cdot \begin{bmatrix} K_R \\ K_G \\ K_B \end{bmatrix} + S \right) \tag{3.12}$$

Assim, verifica-se que a resposta do sensor é função do produto da *DPE* da luz que incide sobre o objecto, pela soma da componente especular (expressa por *S*) com o factor geométrico da reflexão difusa ampliado pela sensibilidade do sensor ao espectro de luz específico do objecto.

De modo similar ao apresentado na definição de invariabilidade à intensidade da iluminação, para que a invariabilidade a brilhos se verifique, é necessário que o espaço de cor efectue o cancelamento de todos os factores à excepção da sensibilidade do sensor ao espectro de luz próprio da matéria do qual o objecto é composto.

Deve-se contudo reiterar que os modelos de cor que verifiquem este tipo de invariabilidade a brilhos podem no entanto ser sensíveis a brilhos originados por condições de iluminação em que o objecto seja atingido com radiação composta apenas por parte do espectro de luz visível, i.e., iluminação colorida. O estudo da invariabilidade a iluminação colorida, apesar de viável, não é considerado no âmbito deste trabalho, embora em determinados cenários (e.g. discotecas e salas de espectáculo) tais condições de iluminação se verifiquem com regularidade.

3.3.3. Espaço de Cor *rgb*

O espaço de cor *rgb*, também conhecido por *RGB normalizado*, resulta de um processo de normalização do espaço de cor *RGB*. Tal processo consiste na quantificação do peso de cada componente de cor, em relação à soma das três componentes primárias (vermelho, verde e azul). Deste modo, a ocorrência de uma variação da luminosidade, que afectará de igual modo as três componentes do espaço de cor *RGB*, não se repercutirá no espaço *rgb*.

Aplicando as transformações especificadas pela equação (3.13) é possível então obter um espaço de cor *rgb*, a partir do espaço *RGB*.

$$\begin{bmatrix} I_r(\mathbf{x}) \\ I_g(\mathbf{x}) \\ I_b(\mathbf{x}) \end{bmatrix} = \frac{255}{I_R(\mathbf{x}) + I_G(\mathbf{x}) + I_B(\mathbf{x})} \cdot \begin{bmatrix} I_R(\mathbf{x}) \\ I_G(\mathbf{x}) \\ I_B(\mathbf{x}) \end{bmatrix} \tag{3.13}$$

Cada componente do novo espaço de cor toma valores compreendidos entre [0, 255]. Um valor nulo numa das suas componentes do espaço *rgb* indica que essa componente não contribui para a cor, enquanto que o valor máximo (255 = 8 *bits*) aponta a componente como única na definição da cor. O plano que define o espaço de cor *rgb* pode ser representado, conforme ilustrado na Figura 3.9, como uma projecção sobre o cubo *RGB*.

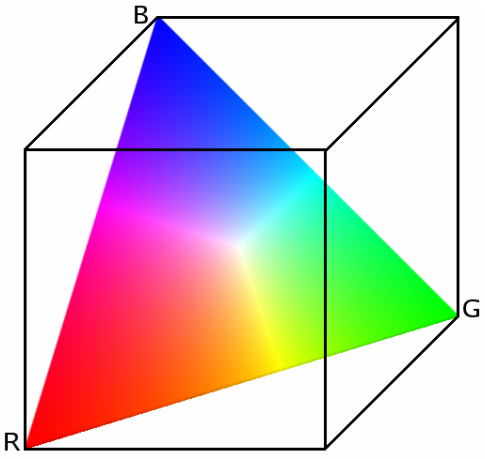


Figura 3.9. Plano *rgb* sobre o cubo *RGB*.

Um exemplo da decomposição de uma imagem colorida, nas três componentes que formam o espaço *rgb*, é apresentado na Figura 3.10. Neste exemplo, as imagens que exibem a contribuição para a cor de cada componente (I_r, I_g e I_b), são apresentadas segundo uma escala de cinzentos onde o branco indica o valor máximo (255) e o preto o valor mínimo admitido pela escala (0).

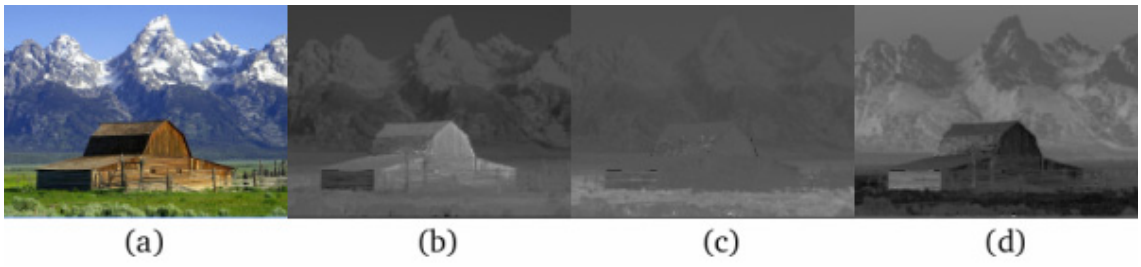


Figura 3.10. (a) Imagem original dividida nas componentes: (b) *r*; (c) *g*; (d) e *b*.

Através da aplicação do modelo de reflexão dicromático é possível aferir analiticamente da invariabilidade de um espaço de cor em relação à intensidade luminosa e a brilhos. As propriedades de invariabilidade do espaço de cor *rgb* a alterações de luminosidade são aqui demonstradas, considerando uma superfície de Lambert iluminada por uma fonte de luz branca (fonte de luz que irradia igual densidade de energia em todos os comprimentos de onda do espectro de luz visível). O estudo da sensibilidade à intensidade da iluminação é realizado através da substituição das componentes de cor *RGB* em (3.13), pelos valores equivalentes definidos por (3.10). Assim, tem-se que:

$$\begin{bmatrix} I_r(\mathbf{x}) \\ I_g(\mathbf{x}) \\ I_b(\mathbf{x}) \end{bmatrix} = \frac{255}{K_R + K_G + K_B} \cdot \begin{bmatrix} K_R \\ K_G \\ K_B \end{bmatrix} \tag{3.14}$$

De acordo com os resultados obtidos em (3.14), verifica-se portanto que o espaço de cor *rgb* não é sensível à direcção e intensidade da iluminação, dependendo apenas da sensibilidade do sensor ao espectro de luz que caracteriza o objecto.

A resposta do espaço de cor a brilhos, é obtida pela resposta do sensor a uma superfície que apresenta uma forte componente especular, e sobre a qual incide uma luz branca. Assim, substituindo as componentes *RGB* das equações (3.13), pelas expressões equivalentes definidas em (3.12), conclui-se que o espaço de cor *rgb* é influenciado pelo coeficiente de reflexão especular, sendo sensível à presença de brilhos na superfície de um objecto, como demonstrado em (3.15).

$$\begin{bmatrix} I_r(\mathbf{x}) \\ I_g(\mathbf{x}) \\ I_b(\mathbf{x}) \end{bmatrix} = \frac{255}{(K_R + K_G + K_B) + 3 \cdot \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}} \cdot \begin{bmatrix} K_R + \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})} \\ K_G + \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})} \\ K_B + \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})} \end{bmatrix} \tag{3.15}$$

3.3.4. Espaço de Cor $c_1c_2c_3$

Um outro modo de representação da cor, ao qual se atribuem propriedades de invariabilidade a determinadas condições de iluminação, é o espaço $c_1c_2c_3$ [Gevers & Smeulders, 1999]. Este modelo de cor é obtido através de uma transformação do espaço *RGB*, determinada pelas seguintes operações:

$$c_1(\mathbf{x}) = \arctan\left(\frac{I_R(\mathbf{x})}{\max(I_G(\mathbf{x}), I_B(\mathbf{x}))}\right) \cdot \frac{255}{90}$$

$$c_2(\mathbf{x}) = \arctan\left(\frac{I_G(\mathbf{x})}{\max(I_R(\mathbf{x}), I_B(\mathbf{x}))}\right) \cdot \frac{255}{90} \qquad (3.16)$$

$$c_3(\mathbf{x}) = \arctan\left(\frac{I_B(\mathbf{x})}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}))}\right) \cdot \frac{255}{90}$$

Onde c_1, c_2 e c_3 tomam valores contidos no intervalo [0, 255] de forma a representar num valor numérico discreto, de 8 *bits*, o intervalo [0º, 90º[. Na Figura 3.11 apresenta-se a decomposição de uma imagem colorida em formato *RGB* de 24 *bits*, nas três componentes do espaço de cor invariante $c_1c_2c_3$. As imagens das componentes deste espaço de cor são apresentadas numa escala de cinzentos.

Figura 3.11. (a) Imagem original dividida nas componentes: (b) c_1; (c) c_2; (d) e c_3.

A análise da sensibilidade do espaço de cor $c_1c_2c_3$ às variações da intensidade da iluminação é realizada pela substituição, nas operações definidas em (3.16), das componentes de cor I_R, I_G, e I_B do espaço *RGB*, pelas suas equivalentes definidas em (3.10). Os resultados desta análise são apresentados em (3.17).

$$c_1(\mathbf{x}) = \arctan\left(\frac{K_R}{\max(K_G, K_B)}\right) \cdot \frac{255}{90}$$

$$c_2(\mathbf{x}) = \arctan\left(\frac{K_G}{\max(K_R, K_B)}\right) \cdot \frac{255}{90} \tag{3.17}$$

$$c_3(\mathbf{x}) = \arctan\left(\frac{K_B}{\max(K_R, K_G)}\right) \cdot \frac{255}{90}$$

Como se pode verificar, as componentes c_1, c_2 e c_3 são unicamente função da sensibilidade do sensor, i.e., K_R, K_G e K_B. Os restantes factores foram cancelados. Por conseguinte, a condição de invariabilidade deste espaço de cor à intensidade luminosa é validada.

Similarmente à análise efectuada no espaço de cor *rgb*, a resposta do espaço $c_1c_2c_3$ a brilhos, é obtida pelo estudo do comportamento do sensor à radiação recebida de uma superfície, com forte componente especular, iluminada por uma fonte de luz branca. Deste modo, substituem-se as componentes de cor *RGB* das equações (3.16), pelas expressões equivalentes apresentadas em (3.12).

$$c_1(\mathbf{x}) = \arctan\left(\frac{K_R + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}{\max(K_G, K_B) + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}\right) \cdot \frac{255}{90}$$

$$c_2(\mathbf{x}) = \arctan\left(\frac{K_G + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}{\max(K_R, K_B) + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}\right) \cdot \frac{255}{90} \tag{3.18}$$

$$c_3(\mathbf{x}) = \arctan\left(\frac{K_B + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}{\max(K_R, K_G) + \frac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}\right) \cdot \frac{255}{90}$$

Analisando as expressões desenvolvidas em (3.18), observa-se que o espaço de cor $c_1c_2c_3$ não só depende da sensibilidade do sensor ao espectro de luz, próprio da matéria do qual o objecto é composto, mas também da componente especular da luz incidente. Conclui-se assim, que o espaço $c_1c_2c_3$ não é invariante a brilhos gerados por fontes de luz branca.

3.3.5. Espaço de Cor $l_1l_2l_3$

O espaço de cor $l_1l_2l_3$ foi desenvolvido por [Gevers & Smeulders, 1999] com a finalidade de determinar a direcção do plano de cor triangular no espaço *RGB*. A sua obtenção, a partir do espaço de cor *RGB*, é especificada em (3.19).

$$\begin{bmatrix} l_1(\mathbf{x}) \\ l_2(\mathbf{x}) \\ l_3(\mathbf{x}) \end{bmatrix} = \frac{255}{(I_R(\mathbf{x})-I_G(\mathbf{x}))^2+(I_R(\mathbf{x})-I_B(\mathbf{x}))^2+(I_G(\mathbf{x})-I_B(\mathbf{x}))^2} \cdot \begin{bmatrix} (I_R(\mathbf{x})-I_G(\mathbf{x}))^2 \\ (I_R(\mathbf{x})-I_B(\mathbf{x}))^2 \\ (I_G(\mathbf{x})-I_B(\mathbf{x}))^2 \end{bmatrix} \quad (3.19)$$

As componentes do espaço de cor $l_1l_2l_3$, tomam valores compreendidos entre [0, 255]. A decomposição de uma imagem colorida, nas três componentes que definem o espaço de cor $l_1l_2l_3$, é ilustrada na Figura 3.12.

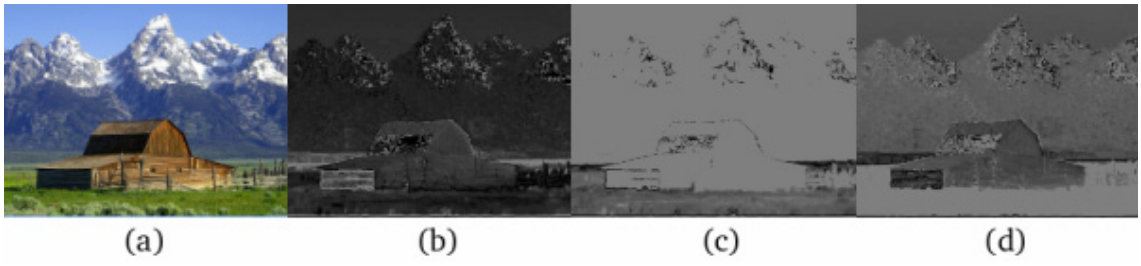


Figura 3.12. (a) Imagem original dividida nas componentes: (b) l_1; (c) l_2; (d) e l_3.

A invariabilidade do espaço de cor $l_1l_2l_3$ à intensidade da iluminação é obtida pela substituição das componentes I_R, I_G, e I_B em (3.19), pelas suas equivalentes definidas por (3.10), da qual resulta:

$$\begin{bmatrix} l_1(\mathbf{x}) \\ l_2(\mathbf{x}) \\ l_3(\mathbf{x}) \end{bmatrix} = \frac{255}{(K_R-K_G)^2+(K_R-K_B)^2+(K_G-K_B)^2} \cdot \begin{bmatrix} (K_R-K_G)^2 \\ (K_R-K_B)^2 \\ (K_G-K_B)^2 \end{bmatrix} \quad (3.20)$$

Os resultados obtidos em (3.20) demonstram que o espaço de cor $l_1l_2l_3$ não é susceptível a perturbações com origem na variação da intensidade da iluminação, uma vez que depende unicamente da sensibilidade do sensor às três componentes do espectro (vermelho, verde e azul).

O estudo da influência de brilhos no espaço de cor $l_1l_2l_3$ é efectuado pela substituição das componentes de cor *RGB* das equações (3.19), pelas expressões equivalentes de (3.12).

Com esta operação, é possível observar que a resposta do sensor a uma superfície com forte componente especular, com origem numa luz branca, é descrita por:

$$\begin{bmatrix} l_1(\mathbf{x}) \\ l_2(\mathbf{x}) \\ l_3(\mathbf{x}) \end{bmatrix} = \frac{255}{(K_R - K_G)^2 + (K_R - K_B)^2 + (K_G - K_B)^2} \cdot \begin{bmatrix} (K_R - K_G)^2 \\ (K_R - K_B)^2 \\ (K_G - K_B)^2 \end{bmatrix} \qquad (3.21)$$

Com base nos resultados obtidos em (3.21), conclui-se que o espaço de cor $l_1l_2l_3$ é também invariante a brilhos. O elemento especular é cancelado, permanecendo apenas função da sensibilidade do sensor (K_R, K_G e K_B).

3.3.6. Espaço de Cor *HSV*

O espaço de cor *HSV* é uma transformação do espaço *RGB* que permite a descrição das cores de um modo mais natural para os humanos. A sigla *HSV* deriva de *Hue* (matriz), *Saturation* (saturação), e *Value* (valor).

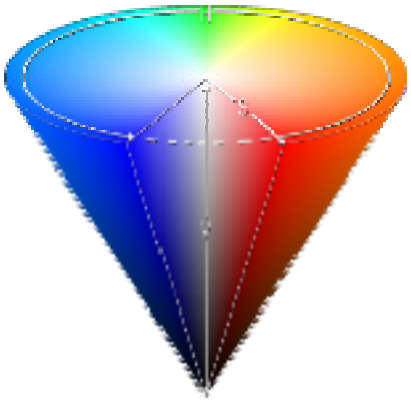


Figura 3.13. Representação cónica do espaço de cor *HSV*.

No espaço de cor *HSV*, a matriz consiste numa medida angular de definição da componente básica da cor. Esta varia de vermelho (para um valor de 0 graus na matriz), passando pelo amarelo (60 graus), verde (120 graus), ciano (180 graus), azul (240 graus) e magenta (300 graus). A componente de saturação descreve a intensidade da cor. Para valores mínimos de saturação obtém-se uma descoloração da cor dominante especificada pela matriz. À medida que o valor da saturação aumenta, aumenta também o nível de coloração da componente matriz. Por fim, a componente de valor permite quantificar a amplitude da cor dominante. Na Figura 3.13 apresenta-se uma representação cónica do espaço de cor *HSV*.

A obtenção de um modelo *HSV* a partir do espaço *RGB* é efectuada através das seguintes operações:

$$I_H(\mathbf{x}) = \begin{cases} \textit{indefinido} \\ \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) \\ 60 \cdot \dfrac{I_G(\mathbf{x}) - I_B(\mathbf{x})}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) - \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))} \\ \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) \geq I_B(\mathbf{x}) \\ 60 \cdot \dfrac{I_G(\mathbf{x}) - I_B(\mathbf{x})}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) - \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))} + 360 \\ \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) < I_B(\mathbf{x}) \\ 60 \cdot \dfrac{I_B(\mathbf{x}) - I_R(\mathbf{x})}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) - \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))} + 120 \\ \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_G(\mathbf{x}) \\ 60 \cdot \dfrac{I_R(\mathbf{x}) - I_G(\mathbf{x})}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) - \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))} + 240 \\ \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_B(\mathbf{x}) \end{cases} \tag{3.22}$$

$$I_S(\mathbf{x}) = \begin{cases} 0 \quad , \textit{ se } \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = 0 \\ \left(1 - \dfrac{\min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))}{\max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x}))}\right) \cdot 255 \quad , \textit{ caso contrário} \end{cases} \tag{3.23}$$

$$I_V(\mathbf{x}) = \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) \tag{3.24}$$

Um exemplo da decomposição do espaço *HSV* em matriz (*H*), saturação (*S*) e valor (*V*) é apresentado na Figura 3.14. Observando a figura, pode-se verificar que na componente matriz se encontram definidas as cores dominantes, enquanto que a saturação indica o nível de coloração, e o valor identifica a sua intensidade.

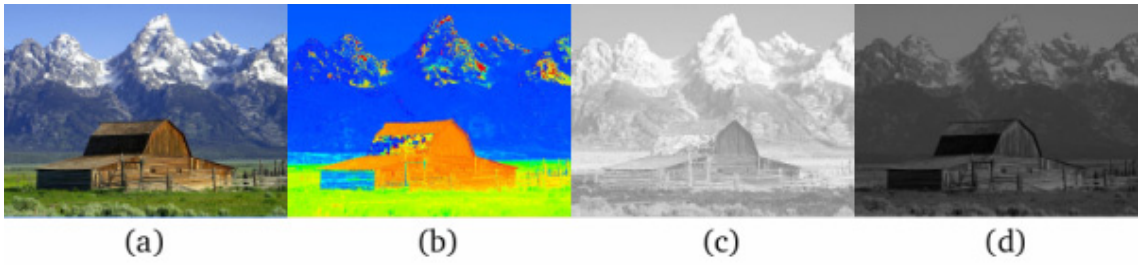


Figura 3.14. (a) Imagem original decomposta em: (b) matriz; (c) saturação; (d) e valor.

A análise da invariabilidade a alterações da iluminação do espaço de cor *HSV* é análoga à efectuada nos restantes modelos de cor. Assim, executa-se a substituição das componentes I_R, I_G, e I_B nas equações (3.22), (3.23) e (3.24), pelas suas equivalentes definidas em (3.10).

$$I_H(\mathbf{x}) = \begin{cases} \textit{indefinido} \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) \\ 60 \cdot \dfrac{K_G - K_B}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) \geq I_B(\mathbf{x}) \\ 60 \cdot \dfrac{K_G - K_B}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 360 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) < I_B(\mathbf{x}) \\ 60 \cdot \dfrac{K_B - K_R}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 120 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_G(\mathbf{x}) \\ 60 \cdot \dfrac{K_R - K_G}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 240 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_B(\mathbf{x}) \end{cases} \tag{3.25}$$

$$I_S(\mathbf{x}) = \left(1 - \frac{\min(K_R, K_G, K_B)}{\max(K_R, K_G, K_B)}\right) \cdot 255 \tag{3.26}$$

$$I_V(\mathbf{x}) = w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}}) \cdot E \cdot \max(K_R, K_G, K_B) \tag{3.27}$$

Verifica-se que as componentes I_H e I_S possuem propriedades de invariabilidade a alterações da luminosidade. Contudo, a componente I_V não apresenta essa propriedade, mostrando-se sensível às condições de iluminação.

O comportamento do espaço de cor *HSV* em relação à presença de brilhos, é efectuado pela substituição das componentes de cor I_R, I_G, e I_B das equações (3.22), (3.23) e (3.24), pelas expressões equivalentes definidas em (3.12).

Analisando os resultados obtidos em (3.28), (3.29) e (3.30), verifica-se que apenas a componente I_H apresenta invariabilidade a brilhos. As restantes componentes (I_S e I_V) são sensíveis à presença de reflexão espectral numa superfície.

$$I_H(\mathbf{x}) = \begin{cases} \textit{indefinido} \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = \min(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) \\ 60 \cdot \dfrac{K_G - K_B}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) \geq I_B(\mathbf{x}) \\ 60 \cdot \dfrac{K_G - K_B}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 360 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_R(\mathbf{x}) \quad \wedge \quad I_G(\mathbf{x}) < I_B(\mathbf{x}) \\ 60 \cdot \dfrac{K_B - K_R}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 120 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_G(\mathbf{x}) \\ 60 \cdot \dfrac{K_R - K_G}{\max(K_R, K_G, K_B) - \min(K_R, K_G, K_B)} + 240 \\ \quad , \; se \quad \max(I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})) = I_B(\mathbf{x}) \end{cases} \tag{3.28}$$

$$I_S(\mathbf{x}) = \left(1 - \frac{\min(K_R, K_G, K_B) + \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}}{\max(K_R, K_G, K_B) + \dfrac{S}{w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}})}} \right) \cdot 255 \tag{3.29}$$

$$I_V(\mathbf{x}) = E \cdot \left(w_d(\vec{\mathbf{n}}, \vec{\mathbf{s}}) \cdot \max(K_R, K_G, K_B) + S \right) \tag{3.30}$$

3.4. Discussão

Os modelos de cor aqui analisados apresentam características que, quando examinadas segundo as necessidades e propósitos deste trabalho, tanto se afiguram como vantagens ou entraves. Interessa sobretudo estudar a resposta dos espaços de cor a três condições: ângulo de visão e geometria do objecto; intensidade da iluminação; e brilhos.

O espaço de cor *RGB* é visto como a base a partir da qual derivam todos os restantes espaços de cor. O modelo *RGB* é sensível a variações do ângulo de visão e geometria dos objectos, oscilações da intensidade da iluminação, e à formação de brilhos nas superfícies. Embora se possa admitir que estes factores constituem importantes limitações em certas tarefas de processamento de imagem, é no entanto possível tirar partido destas características em processos de segmentação de movimento.

A extrema sensibilidade do espaço de cor *RGB* às condições de iluminação permite efectuar uma segmentação de movimento que minimiza os falsos negativos, gerando no entanto um elevado número de falsos positivos (com origem, por exemplo, em sombras e brilhos). Tal desempenho não é de todo prejudicial, desde que associado a uma técnica que realize com sucesso uma segmentação de sombras e brilhos.

Na presença de luz branca, o espaço de cor *rgb* mostrou ser invariante ao ângulo de visão, à geometria do objecto, e a alterações da intensidade luminosa (e.g. sombras). Através do processo de normalização obtém-se um modelo cuja informação relativa às cores se torna imune a alterações da intensidade da iluminação.

Todavia, este modelo é afectado por brilhos e apresenta ainda elevada instabilidade a baixas intensidades de iluminação. Fruto da sua invariabilidade, o espaço de cor *rgb* não permite executar a detecção de regiões afectadas por sombras, desprezado também alguma informação útil para tarefas de segmentação de movimento.

O espaço de cor $c_1c_2c_3$ garante invariabilidade ao ângulo de visão, à geometria do objecto e a variações da intensidade luminosa. A detecção de sombras neste espaço de cor é passível de ser executada utilizando apenas uma única imagem, i.e. sem a necessidade de recorrer a uma imagem de referência. Contudo, tal só é possível graças à aplicação de detectores de contornos (e.g. *Sobel*, *Prewitt* ou *Canny*) sobre os campos de gradientes das componentes do espaço $c_1c_2c_3$ e do espaço *RGB*, distinguindo deste modo os contornos das regiões afectadas por sombras, uma vez que estas não apresentam contornos no espaço invariante de cor. Apesar de se mostrar eficaz, tal operação implica elevados recursos computacionais, podendo ser proibitiva para aplicações com requisitos de tempo-real. Exemplos deste tipo de solução podem ser encontrados nos trabalhos desenvolvidos em [Salvador et al., 2001; Salvador & Ebrahimi, 2002].

O espaço $c_1c_2c_3$ não possui características de invariabilidade a brilhos. Todavia, apresenta vantagens relativamente ao espaço de cor *rgb*. De acordo com o estudo levado a cabo em [Trias, 2005] verifica-se que o espaço $c_1c_2c_3$ garante maior estabilidade que o *rgb*, o qual exibe alguma inconstância em regiões próximas do vértice da origem no cubo *RGB*.

O único modelo de cor, de entre os analisados, que possui invariabilidade a todos os factores considerados é o espaço de cor $l_1l_2l_3$. Não sendo afectado por variações do ângulo de visão, geometria do objecto, oscilações da intensidade luminosa, e brilhos, seria assim de

esperar que este fosse a escolha ideal para tarefas de detecção de movimento. Isto porque idealmente os resultados encontrar-se-iam livres de erros originados por variações das condições de iluminação.

No entanto, em [Salvador, 2004] demonstrou-se que este espaço de cor é mais sensível que os restantes espaços a flutuações dos valores *RGB* originadas por ruído da câmara. O mesmo estudo aponta para a existência de problemas para baixos valores de saturação, i.e., $I_R=I_G=I_B$.

Por fim, o espaço *HSV* apresenta características distintas dos restantes espaços de cor aqui analisados, uma vez que as suas componentes exibem comportamentos divergentes em relação às diferentes condições de iluminação examinadas. Assim, a componente da matriz (*H*), responsável pela definição da cor, mostra-se invariável ao ângulo de visão, à geometria do objecto, às variações da intensidade da iluminação e à presença de brilhos. A saturação, por seu lado, mostra-se sensível aos brilhos reflectidos nas superfícies dos objectos, enquanto que a componente de valor é afectada por todas as condições, i.e. ângulo de visão, geometria do objecto, variação da intensidade da iluminação, e brilhos.

Apesar das características de invariabilidade da componente de matriz, esta exibe instabilidade em regiões próximas do eixo que define a escala de cinzentos (baixos valores de saturação) [Salvador, 2004]. Este problema afecta também a componente de saturação. A componente de matriz tem ainda como inconveniente a sua reduzida capacidade de discriminação, que pode causar importantes limitações em técnicas de segmentação de movimento.

Na Tabela 3.1 resume-se o desempenho dos cinco espaços de cor analisados em relação a: alteração do ângulo de visão e geometria do objecto; variação da intensidade da iluminação; e tolerância a brilhos. O símbolo "×" identifica uma relação de invariabilidade.

Tabela 3.1. Quadro de características de invariabilidade dos espaços de cor analisados.

	RGB	***rgb***	$c_1c_2c_3$	$l_1l_2l_3$	***H***	***S***	***V***
Ângulo de Visão e Geometria		×	×	×	×	×	
Intensidade da Iluminação		×	×	×	×	×	
Brilhos				×	×		

Os problemas de instabilidade e sensibilidade ao ruído dos espaços de cor (*rgb*, $c_1c_2c_3$, $l_1l_2l_3$ e *HSV*) são sintetizados pela Tabela 3.2. Note-se que tais informações serão relevantes na escolha dos modelos de representação de cor a ser utilizados nos processos de segmentação e seguimento de objectos.

Tabela 3.2. Instabilidade e sensibilidade dos espaços de cor a condições de baixa intensidade luminosa e saturação.

		rgb	$c_1c_2c_3$	$l_1l_2l_3$	***H***	***S***	***V***
Instabilidade	$I_R=I_G=I_B=0$	×	×	×	×	×	
	$I_R=I_G=I_B\neq0$			×	×		
Sensibilidade	$I_R=I_G=I_B=0$	×	×	×	×	×	
	$I_R=I_G=I_B\neq0$			×	×		

Capítulo 4

4. Segmentação de Objectos em Movimento

No corrente capítulo são descritas as técnicas utilizadas para a segmentação de objectos em movimento a partir de sequências de imagens digitalizadas e não comprimidas, obtidas por uma câmara monocular fixa. As técnicas aqui propostas possibilitam a extracção de regiões de pontos de uma imagem, relativos a objectos não pertencentes ao plano de fundo, livres de erros referentes a sombras, brilhos e fantasmas.

A estratégia adoptada permite realizar a segmentação de objectos em movimento por subtracção de plano de fundo adaptativo, sem a necessidade da especificação de características da câmara de vídeo, condições de iluminação ou da contextualização da cena.

Principia-se este capítulo com a definição do problema da segmentação. Prossegue-se com o estudo de técnicas para a detecção de sombras e brilhos, explorando diversos espaços de cor. Seguidamente, apresenta-se o método proposto para a identificação de fantasmas originados por técnicas de segmentação baseadas na subtracção por plano de fundo. O capítulo termina com uma discussão da solução proposta, que é precedida da avaliação das suas capacidades.

4.1. Definição do Problema da Segmentação

A tarefa da segmentação de imagens digitalizadas consiste na decomposição de uma imagem em várias regiões ou classes. Este processo implica que cada ponto da imagem seja atribuído a uma determinada região, de tal modo que, da reunião dos pontos contidos pela totalidade das regiões, é possível reconstituir na íntegra a imagem processada. A segmentação pode então ser formalmente definida do seguinte modo:

Seja E o conjunto finito de pontos de uma imagem I, e seja $P(\,)$ o predicado lógico de homogeneidade. Então a segmentação do conjunto E consiste na sua decomposição em vários subconjuntos (ou regiões) R_i, i=1,…,n com $n \in IN$, tal que,

$$E = \bigcup_{i=1}^{n} R_i \quad , \text{com} \quad R_i \cap R_j = \varnothing \quad , \forall i \neq j$$
$$P(R_i) = \text{Verdadeiro} \quad , \forall i = 1,...,n$$
$$P(R_i \cup R_j) = \text{Falso} \quad , \forall i \neq j \quad , R_i \quad \text{é adjacente a} \quad R_j$$

sendo que R_i e R_j são adjacentes, se existirem dois pontos adjacentes $z, w \in E$, com $z \in R_i$ e $w \in R_j$.

A segmentação é utilizada em diversos domínios de aplicação na área da visão por computador. Em determinados contextos, como a pesquisa de conteúdos em bases de dados de imagens, a segmentação está associada à **detecção das regiões homogéneas de uma única imagem**, que identificam os vários objectos em cena, quer seja pela cor que apresentam [Lucchese & Mitra, 1998], pela textura que os caracterizam [Muneeswaran et al., 2006], ou através de uma combinação dos dois tipos de informação [Chen et al., 2005].

Por outro lado, em aplicações de vídeo-vigilância, a segmentação consiste na detecção de objectos em movimento em sequências de imagens de vídeo [Zhang & Lu, 2001]. Em aplicações de **segmentação de movimento**, a classificação dos pontos da imagem é efectuada em uma de duas classes possíveis: **plano de fundo** e **movimento**. No entanto, cada região classificada como movimento pode conter um ou vários objectos. A correcta atribuição dos pontos segmentados aos objectos correspondentes não é da responsabilidade do processo de segmentação, mas sim do posterior processo de seguimento de objectos.

Na Figura 4.1 ilustram-se os dois tipos de segmentação, i.e., (c) segmentação de regiões homogéneas da imagem (a), e (d) segmentação de movimento pela subtracção da imagem (a) à imagem de referência (b).

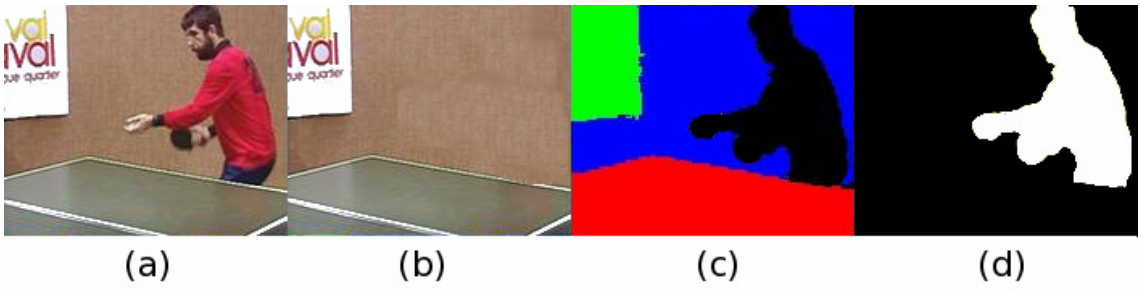


Figura 4.1. (a) Imagem alvo; (b) Referência utilizada na segmentação de movimento por subtracção de plano de fundo; (c) Segmentação de regiões homogéneas; (d) Segmentação de movimento.

A segmentação de regiões afectadas por movimento, em imagens de uma sequência de vídeo, é uma tarefa fundamental e crítica na identificação de eventos anormais ou passíveis de originar uma quebra de segurança. É pela análise dos resultados produzidos pela segmentação que se torna viável monitorizar a actividade humana através do seguimento de pessoas e veículos num espaço.

Um método de segmentação de movimento ideal deveria ser suficientemente robusto contra alterações ambientais, mas sensível o bastante para detectar os objectos de interesse em movimento. Todavia, este tipo de segmentação em ambientes reais não é perfeito, debatendo-se com dificuldades que se devem sobretudo à natureza da técnica utilizada no processamento da estimativa das regiões afectadas por movimento, bem como do tipo de perturbações que afectam os seus resultados.

De acordo com as considerações apresentadas na Secção 2.3.1, adopta-se neste trabalho uma abordagem de segmentação de movimento por subtracção de plano de fundo adaptativo, por ser a técnica que melhor se adapta à aplicação em questão, em particular devido ao seu desempenho. Este tipo de técnica mostra-se contudo sensível a fenómenos, tais como: a variação de luminosidade ao longo do dia ou devido à oclusão da luz solar por nuvens; condições atmosféricas adversas como o nevoeiro, a chuva, a neve e o vento; bem como às sombras dos objectos em movimento. Um outro problema associado às técnicas de segmentação de movimento por subtracção do plano de fundo tem a ver com a manutenção da imagem de referência e das condicionantes que habitualmente envolvem a sua inicialização, que frequentemente originam a segmentação de fantasmas.

4.2. Detecção de Sombras e Brilhos

As condições de iluminação originam habitualmente perturbações em diversos algoritmos de visão por computador. Em especial a presença de diferentes níveis de iluminação numa cena, originando distribuições não uniformes da luz; a variabilidade da intensidade luminosa ao longo do tempo; e a posição da fonte de luz relativamente ao eixo definido pelo ponto de observação e a área objecto de interesse, bem como a deslocação da fonte de luz ao longo do tempo. Tudo isto são factores a considerar aquando da implementação de um qualquer sistema de visão por computador.

4.2.1. Sombreamento

A distribuição não uniforme da intensidade luminosa pelas superfícies dos objectos nos quais incide directamente uma fonte de luz é chamada de **sombreamento**[16] [Fukusima, 1997]. Apesar de ser considerado como um indício de profundidade relativamente ténue, o sombreamento é de extrema importância na percepção do espaço tridimensional. Tal facto tem vindo a ser explorado no desenvolvimento de sistemas gráficos em visão cibernética, para o auxílio da percepção do espaço tridimensional. Apesar da natureza dos sistemas de vídeo-vigilância consistir na monitorização de espaços tridimensionais, não existe uma necessidade vital de percepcionar a profundidade dos objectos que integram o ambiente sobre observação. Os alvos de interesse são os objectos em movimento, cujos atributos identificativos e caracterizadores não carecem de qualquer informação sobre o sombreamento.

Figura 4.2. O sombreamento observado num túnel.

[16] Tradução adoptada do termo inglês: *shadowing*.

4.2.2. Sombras Próprias e Projectadas

As **sombras** resultam da projecção do bloqueio da luz, por um objecto ou obstáculo, sobre as superfícies adjacentes. Numa imagem, as sombras podem ser classificadas em duas categorias: **sombras próprias** e **sombras projectadas** [Knill et al., 1997; Mamassian et al., 1998]. Quando as sombras incidem sobre a superfície do próprio objecto denominam-se sombras próprias. Por outro lado, designa-se por sombra projectada toda a área abrangida pela oclusão à fonte de luz por um outro objecto. Este último tipo de sombra pode-se ainda decompor em sombras projectadas estáticas, caso sejam causadas por objectos estáticos (e.g. edifícios e árvores), e sombras projectadas dinâmicas, se resultarem de objectos em movimento (e.g. pessoas e veículos). A Figura 4.3 ilustra os dois géneros de sombras projectadas.

No contexto da vídeo-vigilância, as sombras projectadas estáticas consideram-se pertencentes ao ambiente a monitorizar, uma vez que não se alteram em curtos períodos de tempo. De modo similar, as sombras próprias de objectos estáticos também não possuem qualquer relevância em aplicações de segmentação de objectos em movimento. As sombras próprias pertencentes a objectos em movimento não requerem igualmente cuidados especiais, uma vez que fazem parte integrante dos objectos a segmentar. Por outro lado, as sombras projectadas dinâmicas têm uma influência negativa na segmentação de objectos em movimento. Devido ao facto de estas acompanharem a deslocação dos objectos, provocando alterações significativas na intensidade luminosa nas zonas onde são projectadas, as sombras projectadas dinâmicas são frequentemente segmentadas juntamente com os objectos, alterando assim a sua forma e dimensões.

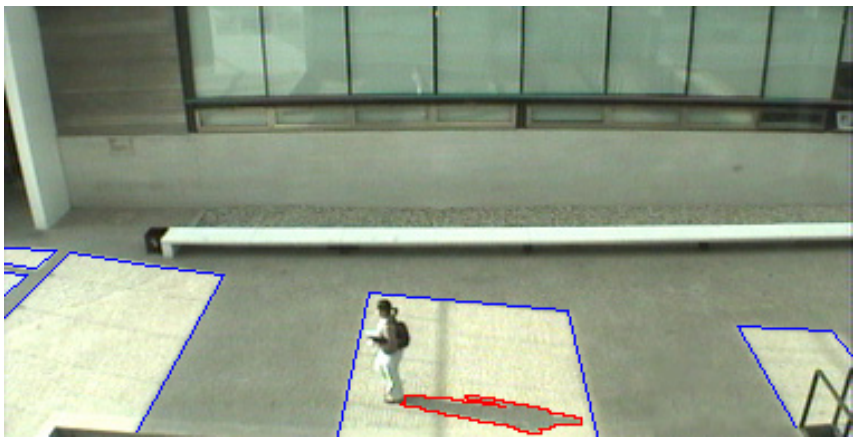


Figura 4.3. Exemplos de: sombras projectadas estáticas (contornos a azul) e sombras projectadas dinâmicas (contornos a vermelho).

4.2.3. Brilhos

Um outro factor com influência na segmentação de movimento é o **brilho**. Este fenómeno é patente sempre que se verifique o aumento da intensidade da radiação luminosa incidente sobre um corpo que possua forte coeficiente de reflexão especular, como ilustrado na Figura 4.4. Neste caso, e considerando que o coeficiente de reflexão especular do corpo é aproximadamente constante para o espectro de luz visível, o sinal recebido pelo sensor de aquisição de imagem é afectado pela adição desta componente, com propriedades de radiação espectral semelhantes à fonte de luz.

Os brilhos influenciam assim, de forma significativa, a percepção da cor dos objectos em determinados modelos de representação de cor, constituindo uma perturbação relevante em qualquer sistema de segmentação de movimento. Erros de segmentação originados por brilhos exibidos nas superfícies de objectos estáticos pertencentes à cena são frequentes devido à errada interpretação do fenómeno em certos espaços de cor.

Figura 4.4. Brilhos no *capot* do automóvel, resultantes da reflexão especular.

4.2.4. Cancelamento da Influência das Condições de Iluminação

A resolução dos problemas associados a sombras e brilhos não é trivial, tendo vindo a merecer a atenção da comunidade científica no sentido de o obviar. Apesar da natural diversidade de metodologias propostas, decorrente de numerosas abordagens ao problema sobre os mais diversos domínios de aplicação (e.g. na detecção e correcção de sombras em imagens estáticas adquiridas via satélite [Suzuki et al., 2000; Polidorio et al., 2003; Sarabandi et al., 2004], na vídeo-vigilância rodoviária [Yoneyama et al., 2003; Matsushita et al., 2004],

e em sequências de vídeo para vigilância de espaços públicos [Xu & Ellis, 2001; Salvador et al., 2003]), verifica-se que dos esforços nesse sentido resultam principalmente em três tipos de abordagens ao problema: o **controlo das condições de iluminação** sobre o ambiente a monitorizar; a utilização de **modelos geométricos** na detecção de sombras; e a definição de **modelos invariantes de representação de cor**.

Apesar de comum em aplicações industriais, o **controlo das condições de iluminação** [Xu & Zhang, 2001] não é exequível no âmbito da vídeo-vigilância. Num ambiente industrial, o alvo de interesse encontra-se próximo da câmara de vídeo, inserido num ambiente isolado e por conseguinte, passível de ser controlado. Contrariamente, em aplicações de vídeo-vigilância os objectos a monitorizar encontram-se consideravelmente afastados do ponto de observação, podendo estes deslocarem-se por uma ampla área, usualmente em ambientes externos ou em ambientes permissivos a influências externas, nos quais se torna impraticável qualquer tentativa de controlo das condições de iluminação.

Os **modelos geométricos** na detecção de sombras tiram partido do conhecimento prévio da forma e sentido de deslocação dos objectos, bem como da estrutura rígida e indeformável que estes apresentam. Este tipo de abordagem encontrou aplicação em áreas como a monitorização e seguimento de veículos [Funka-Lea & Bajcsy, 1995; Yoneyama et al., 2003]. A transposição desta técnica para a vigilância de locais públicos, cujo alvo de interesse engloba pessoas e veículos, não é porém viável. Os seres humanos podem exibir um elevado número de formas durante deslocamentos, fruto das características biomecânicas dos seus corpos. Tal facto impossibilita a definição de um modelo geométrico simples, adequado à sua caracterização.

A cor é um fenómeno de percepção relacionado com a resposta do olho humano a diferentes comprimentos de onda do espectro electromagnético visível. Na maioria dos casos, a cor é descrita como uma combinação de três cores primárias (e.g. vermelho, verde e azul para o espaço de cor *RGB*). Apesar de em espaços de cor como o *RGB* a variação da intensidade luminosa provocar profundas alterações na definição da cor, os **modelos invariantes de representação de cor** possibilitam a conservação da informação da cor, independentemente do nível de iluminação, indo deste modo ao encontro das necessidades específicas dos sistemas de vídeo-vigilância. Tal como foi descrito anteriormente, vários modelos de cor foram propostos: *rgb*, $c_1c_2c_3$, $l_1l_2l_{33}$ e *HSV*.

4.2.5. Selecção do Espaço Invariante de Representação de Cor

A análise realizada no terceiro capítulo desta tese, aos espaços invariantes de representação de cor revelou características distintas para cada modelo. Os espaços de cor *rgb* e $c_1c_2c_3$ demonstram não ser sensíveis a sombras, mas não manifestam invariabilidade a brilhos. Apenas os modelos $l_1l_2l_3$ e *HSV* possuem características de invariabilidade a sombras e brilhos, posicionando-se como os mais aptos para a detecção destes fenómenos.

O espaço de cor $l_1l_2l_3$ apresenta sérios problemas no eixo que define a gama de valores da escala de cinzentos no cubo *RGB*. Fruto do processo de transformação que dá origem a este modelo, torna-se impossível distinguir, por exemplo, um ponto que exiba a cor branca de outro com coloração próxima ao preto absoluto. O espaço *HSV*, apesar da sua componente de matriz (*H*) exibir o mesmo problema, apresenta no entanto especificidades que permitem retirar maior proveito relativamente ao espaço de cor $l_1l_2l_3$.

A componente de matriz do espaço *HS*V é invariável a sombras e brilhos. Por outro lado, a saturação mostra-se vulnerável à presença de brilhos, permanecendo no entanto invariável a sombras. Sensível a todas as alterações induzidas sobre as condições de iluminação encontra-se a componente de valor. Esta componente é de enorme relevância, uma vez que não sofre de problemas de instabilidade ou de sensibilidade ao ruído. A conjugação das propriedades distintas das três componentes deste espaço de cor, possibilita uma descrição intuitiva das cores [Herodotou et al., 1998] e de como as sombras e os brilhos se formam.

De acordo com [Cucchiara et al., 2001b], as sombras projectadas sobre um plano de fundo não produzem variações significativas na matriz. Porém, observa-se uma diminuição da saturação e da componente de valor. Pelo contrário, no caso dos brilhos verifica-se um acréscimo da componente de valor. Tirando partido destes factos, é exequível a detecção de sombras e brilhos por comparação da alteração da luminosidade entre uma imagem de referência e a imagem alvo, necessitando para tal que a codificação das imagens seja efectuada para o espaço de cor *HSV*.

O modelo de representação de cor *HSV* goza assim de propriedades que permitem executar a detecção e segmentação de sombras e brilhos de uma forma intuitiva e com moderados recursos computacionais. Como tal optou-se, neste trabalho, por realizar a detecção de sombras e brilhos recorrendo a este espaço de cor.

4.2.6. Técnica de Segmentação de Sombras e Brilhos

Para efectuar a detecção de sombras e brilhos, propõe-se uma abordagem que utiliza a imagem alvo de análise (*I*), e uma imagem de referência (*B*) que define o plano de fundo. Ambas as imagens são representadas pelo espaço de cor *HSV*. Deste modo, o primeiro passo na segmentação de sombras e brilhos consiste na conversão para este modelo de cor, de todas as imagens adquiridas, a partir do espaço *RGB* (formato disponibilizado pelo sistema de aquisição de imagem). A transformação de *RGB* para *HSV* é realizada através das equações (3.22), (3.23) e (3.24).

Após a obtenção do espaço de cor seleccionado para o efeito, é então necessário definir um processo que possibilite a correcta detecção de sombras e brilhos. No espaço *HSV*, a matriz mantém-se inalterável ou experimenta apenas pequenas variações quando a superfície de um objecto é sujeitada a uma alteração da intensidade luminosa. Explorando estas características invariantes, é possível identificar regiões de pontos de uma imagem, afectadas por sombras, como demonstrado em [Cucchiara et al., 2002].

Tendo por base este conceito, é possível aperfeiçoar a técnica de modo a segmentar não só as sombras, como também os brilhos induzidos por outras fontes de luz branca (e.g. lanternas e faróis de veículos) ou reflexões (provocadas por janelas e espelhos) não incorporadas na imagem de referência. Assim, no âmbito deste trabalho, foi desenvolvida uma nova técnica para a detecção de sombras e brilhos.

Para tal, considere-se que $I_H(\mathbf{x})$, $I_S(\mathbf{x})$ e $I_V(\mathbf{x})$ representam respectivamente as componentes matriz, saturação e valor na coordenada $\mathbf{x}=(x, y)$ da imagem alvo. Analogamente, $B_H(\mathbf{x})$, $B_S(\mathbf{x})$ e $B_V(\mathbf{x})$ representam as três componentes do espaço *HSV*, na coordenada $\mathbf{x}=(x, y)$ da imagem de referência *B*.

De acordo com a análise efectuada ao espaço de cor *HSV*, na Secção 3.3.6, verifica-se então que a componente de matriz não é afectada por variações da intensidade da iluminação ou por brilhos provocados por uma fonte de luz branca. Por outro lado, a saturação é sensível a brilhos, enquanto que a componente de valor é sensível a todos os factores. Deste modo, a detecção de sombra num ponto, realizada pela comparação com uma imagem de referência, pode ser modelada matematicamente pela seguinte condição:

$$I_V^t(\mathbf{x}) < B_V^t(\mathbf{x}) \quad \wedge \quad I_S^t(\mathbf{x}) = B_S^t(\mathbf{x}) \quad \wedge \quad I_H^t(\mathbf{x}) = B_H^t(\mathbf{x}) \tag{4.1}$$

Assim, segundo a condição descrita em (4.1), um ponto na coordenada $\mathbf{x}=(x, y)$ da imagem *I* é classificado como sombra se a componente de valor for inferior ao definido para a mesma coordenada da imagem de referência *B*, e se as componentes de saturação e matriz se mantiverem inalteradas.

Explorando as propriedades deste espaço de cor, a detecção de brilhos pode então ser descrita como um aumento da componente de valor, da diminuição da saturação, mantendo no entanto a componente de matriz. Ou seja:

$$I_V^t(\mathbf{x}) > B_V^t(\mathbf{x}) \quad \wedge \quad I_S^t(\mathbf{x}) < B_S^t(\mathbf{x}) \quad \wedge \quad I_H^t(\mathbf{x}) = B_H^t(\mathbf{x}) \tag{4.2}$$

Esta abordagem é contudo demasiado restritiva, não permitindo flutuações próprias de ambientes reais. Tais variações devem-se sobretudo à existência de inter-reflexões, i.e. radiações provenientes de reflexões das superfícies de objectos e que incidem sobre a superfície de um outro objecto. Como essas reflexões espelham diferentes gamas do espectro de luz, características de cada corpo, a sua influência na *DPE* que ilumina a superfície do objecto provoca uma alteração da percepção da cor.

A natureza do espectro irradiado pela fonte de luz é igualmente uma origem de perturbações. Na realidade, tanto a radiação solar como a obtida artificialmente por meio de lâmpadas, não satisfazem plenamente a definição de luz branca. Isto porque não se verifica uma distribuição de potência similar para todos os comprimentos de onda do espectro de luz visível. Na verdade, no que diz respeito à radiação solar, a *DPE* varia ao longo do dia, como demonstrado pela Figura 4.5.

Segundo o padrão *CIE*, a série D especifica as várias distribuições do espectro da luz solar, que se manifestam ao longo do dia. Assim, a iluminação característica do nascer ou pôr-do-sol é definida pela temperatura de cor de 5000º Kelvin (D50). À iluminação própria do meio da manhã ou meio da tarde corresponde uma temperatura de 5500º Kelvin (D55). Já a iluminação que se observa ao meio-dia é representada por uma temperatura de cor de 6500º Kelvin (D65).

A *CIE* identificou também as componentes do espectro de iluminação artificial como as lâmpadas de filamento (série A) e lâmpadas fluorescentes (série F).

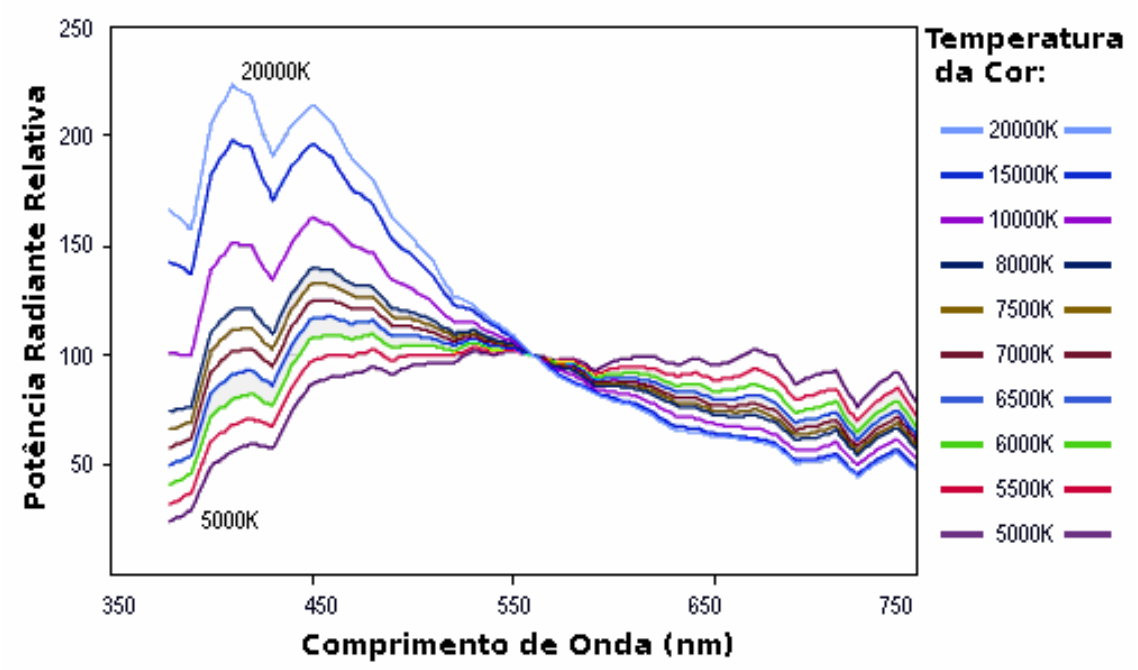


Figura 4.5. Padrões de iluminação definidos pela *Commission Internationale de L'éclairage* (*CIE*).

De modo a avaliar a resposta a situações reais, das condições definidas em (4.1) e (4.2), recorreu-se a imagens capturadas pela câmara número 1, do grupo de teste, do primeiro conjunto de imagens propostas para o *Performance Evaluation of Tracking and Surveillance Workshop* [PETS, 2001]. Deste conjunto de teste, utilizou-se a imagem número 1 como referência, Figura 4.6 (a), e a imagem número 872 como alvo, Figura 4.6 (b), onde se verifica a presença de dois veículos e de quatro pessoas. Deve-se no entanto fazer notar que as imagens utilizadas se encontravam comprimidas em formato *JPEG*, pelo que será de esperar a presença de ruído resultante do processo de compressão.

Os resultados da segmentação de movimento, sombras e brilhos são apresentados pela Figura 4.6 (c). Nessa imagem, os pontos a verde indicam a presença de brilhos, enquanto que as sombras são assinaladas a vermelho. As regiões de pontos que apresentam uma coloração azul especificam zonas da imagem onde se detectou movimento, sendo que considera-se que um ponto da imagem é afectado por movimento se se verificar a seguinte condição no espaço de cor *RGB*:

$$\left|I_R(\mathbf{x}) - B_R(\mathbf{x})\right| > T \quad \wedge \quad \left|I_G(\mathbf{x}) - B_G(\mathbf{x})\right| > T \quad \wedge \quad \left|I_B(\mathbf{x}) - B_B(\mathbf{x})\right| > T \qquad (4.3)$$

com T=16, para uma escala de valores entre [0, 255].

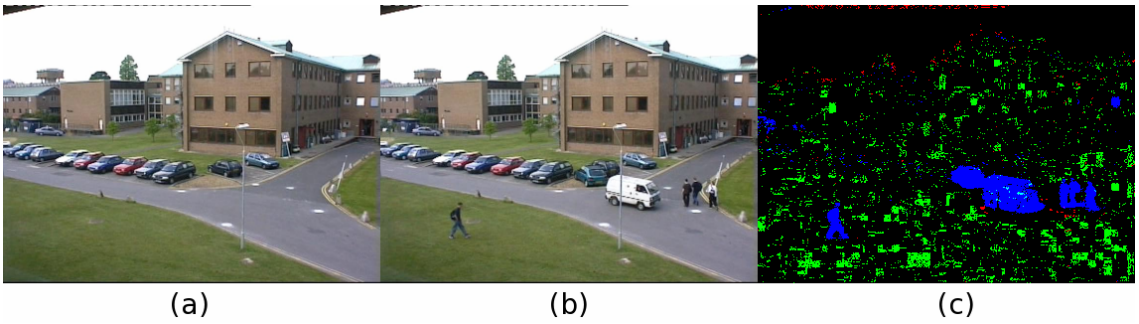


Figura 4.6. Teste a sombras e brilhos, utilizando as condições definidas em (4.1) e (4.2).

Pela análise dos resultados da segmentação apresentados na Figura 4.6 (c), verifica-se que os objectos não pertencentes à imagem de referência são segmentados na sua totalidade. Contudo, as sombras projectadas por esses objectos são também consideradas como movimento. Este facto origina a ocorrência de erros na segmentação, como se pode observar pela segmentação do grupo constituído por três pessoas, onde estas se encontram conectadas por pontos classificados erradamente como movimento.

A detecção de sombras apenas se verificou num número reduzido de pontos isolados, sendo o ruído a sua causa mais provável. Já no que respeita a brilhos, foram detectadas várias regiões da imagem afectadas por este fenómeno. Apesar de não se esperar à partida tal abundância de zonas de brilho, a sua aparição pode ser explicada pela considerável diferença temporal que separa a imagem de referência da imagem alvo. De facto, analisando a luminosidade das duas imagens, verifica-se que a imagem alvo exibe maior índice deste factor que a imagem de referência, para os objectos pertencentes ao plano de fundo. Tal facto, não é no entanto perceptível a olho nu.

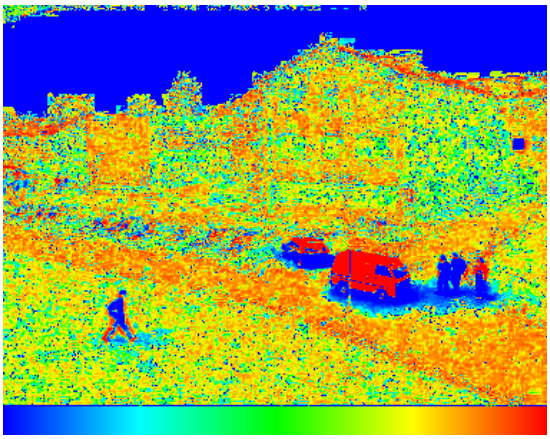


Figura 4.7. Variação, em escala logarítmica, da luminosidade entre a imagem de referência e a imagem alvo. O verde indica luminosidade semelhante, o azul indica uma diminuição, e o vermelho um aumento em relação à imagem de referência.

Constata-se então a necessidade de admitir alguma tolerância nas condições (4.1) e (4.2). Por conseguinte, no que diz respeito a sombras e brilhos, deve-se permitir uma variação da componente de matriz em ambos os sentidos, i.e. aumento ou diminuição do comprimento de onda dominante. Contudo, é necessário estabelecer o limite máximo permitido para tal variação. Como tal, dado que a matriz se encontra dividida em seis cores básicas (vermelho, amarelo, verde, ciano, azul e magenta) num intervalo de 360º, então a sua variação não deverá ultrapassar os 60º (360º / 6 cores = 60º por cor), ou seja aproximadamente 16% de uma escala de [0, 255].

$$\left| I_H^t(\mathbf{x}) - B_H^t(\mathbf{x}) \right| \leq \tau_H \tag{4.4}$$

Note-se que τ_H admite como valor máximo 60º ou 16% da escala de discretização da matriz.

Idealmente, a saturação mantém-se inalterada na presença de sombras, diminuindo com a ocorrência de brilhos, como se comprovou na Secção 3.3.6. Todavia, quando o ponto de referência apresenta uma coloração próxima à gama definida pelo eixo diagonal do cubo *RGB* que define a escala de cinzentos, a resposta da componente de saturação tende a tornar-se instável. Nesta gama de valores, pequenas oscilações induzidas por ruído do sensor de aquisição de imagem ou outro tipo de perturbações podem provocar um inesperado aumento ou uma diminuição fortuita da saturação. Por este motivo, ao invés de se considerar uma saturação constante para sombras ou uma diminuição na presença de brilhos, passa-se a quantificar a sua variação absoluta. Porém, esta variação deve estar contida dentro de um determinado limite, de modo a que os brilhos sejam correctamente discriminados das oscilações provocadas por movimentos de objectos em cena.

A natureza do espectro da fonte de luz influencia também a componente de saturação em superfícies com forte componente especular, como se pode constatar pelas equações (3.29), (3.11) e (3.5). Contudo, não é viável definir analiticamente o limite máximo permitido para a sua variação (τ_S), pelo que este parâmetro tem necessariamente que ser definido empiricamente.

$$\left| I_S^t(\mathbf{x}) - B_S^t(\mathbf{x}) \right| \leq \tau_S \tag{4.5}$$

As condições definidas para a componente de valor em (4.1) e (4.2) devem ser igualmente redefinidas, de modo a possibilitar uma maior aproximação aos efeitos observados em

cenários reais. A ideia de que a ocorrência de sombras sobre a superfície de um objecto origina uma diminuição da componente de valor sem se considerar a magnitude dessa variação, é demasiado simplista. Do mesmo modo, a ocorrência de brilhos não pode ser somente descrita como tendo um efeito de amplificação da componente de valor. É necessário definir uma taxa de variação permitida a sombras e brilhos, acima da qual o fenómeno passe a ser classificado como movimento. Um factor que possibilite estabelecer um limite mínimo de variação, próprio a sombras e brilhos, deve também ser adicionado de forma a atenuar pequenas flutuações dessa componente.

$$\alpha \leq \frac{I_V^t(\mathbf{x})}{B_V^t(\mathbf{x})} \leq \beta \tag{4.6}$$

$$\frac{1}{\beta} \leq \frac{I_V^t(\mathbf{x})}{B_V^t(\mathbf{x})} \leq \frac{1}{\alpha} \tag{4.7}$$

Através do cálculo da razão da componente de valor num ponto da imagem alvo, pela componente de valor especificada na mesma coordenada da imagem de referência, obtém-se uma medida de magnitude da variação, em relação ao valor de referência. Assim, no caso da detecção de sombras (4.6), a taxa de variação permitida é definida por α, enquanto que β especifica o limite mínimo da variação. Na presença de brilhos, a gama de valores permitidos para a variação da magnitude ocorre no sentido inverso (4.7).

Adicionando as referidas tolerâncias às condições definidas em (4.1) e (4.2), é possível descrever matematicamente a geração de uma máscara de sombras (*MS*) e de uma máscara de brilhos (*ML*) através das seguintes expressões:

$$MS^t(\mathbf{x}) = \begin{cases} 1 & ,se \quad \alpha \leq \dfrac{I_V^t(\mathbf{x})}{B_V^t(\mathbf{x})} \leq \beta \quad \wedge \quad \left|I_S^t(\mathbf{x}) - B_S^t(\mathbf{x})\right| \leq \tau_S \quad \wedge \quad \left|I_H^t(\mathbf{x}) - B_H^t(\mathbf{x})\right| \leq \tau_H \\ 0 & ,caso \quad contrário \end{cases} \tag{4.8}$$

$$ML^t(\mathbf{x}) = \begin{cases} 1 & ,se \quad \dfrac{1}{\beta} \leq \dfrac{I_V^t(\mathbf{x})}{B_V^t(\mathbf{x})} \leq \dfrac{1}{\alpha} \quad \wedge \quad \left|I_S^t(\mathbf{x}) - B_S^t(\mathbf{x})\right| \leq \tau_S \quad \wedge \quad \left|I_H^t(\mathbf{x}) - B_H^t(\mathbf{x})\right| \leq \tau_H \\ 0 & ,caso \quad contrário \end{cases} \tag{4.9}$$

Porém, é ainda necessário definir os valores dos parâmetros: α, β e τ_S. Cucchiara e seus colaboradores em [Cucchiara et al., 2001] propuseram um conjunto de valores para a segmentação de sombras, sendo que os mesmos podem ser utilizados na detecção de

brilhos. Assim, o limite máximo para a variação da saturação é de 15% da sua escala de discretização. O parâmetro α toma valores compreendidos entre [0.75, 0.85], enquanto que β admite valores do intervalo [0.90, 0.97].

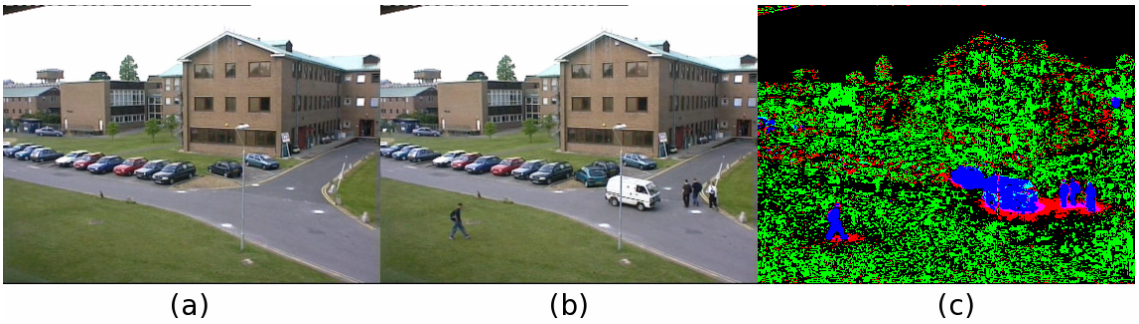


Figura 4.8. Teste a sombras e brilhos, utilizando as condições definidas em (4.8) e (4.9).

Pela análise dos resultados ilustrados na Figura 4.8, verifica-se que as condições (4.8) e (4.9) asseguram uma eficaz segmentação de sombras e brilhos. No que diz respeito à segmentação de sombras a melhoria é notória, como se pode verificar pela Figura 4.8 (c) onde as regiões representadas a vermelho definem correctamente as sombras projectadas pelos objectos em movimento.

4.3. Segmentação de Movimento

A fusão de uma técnica de segmentação de movimento baseada na subtracção do plano de fundo adaptativo, com a capacidade de detecção de regiões da imagem afectadas por sombras e brilhos, permite obter resultados de segmentação com precisão superior.

O método mais elementar para a segmentação de movimento por subtracção de plano de fundo consiste em analisar, para cada ponto da imagem, a variação da informação de cor entre uma imagem de referência e a imagem alvo. Apesar de existirem várias formas de executar essa medição (e.g. análise da média das componentes de cor ou o cálculo da distância euclidiana das componentes), um processo que permite obter um maior detalhe na medição dessa variação, sem necessidade de treino, baseia-se na análise independente da variação de cada componente de cor. A classificação de movimento pode então ser representada pela geração da máscara primária de movimento (*MPM*), definida em (4.10).

$$MPM^t(\mathbf{x}) = \begin{cases} 1 & , \text{ se } \arg\max_{R,G,B} \left| I^t_{R,G,B}(\mathbf{x}) - B^t_{R,G,B}(\mathbf{x}) \right| > T \\ 0 & , \text{ caso contrário} \end{cases} \tag{4.10}$$

Assim, consideram-se em movimento todos os pontos para os quais exista uma componente de cor cuja diferença absoluta, entre a imagem de referência e a imagem alvo, seja superior a um determinado valor (*T*). Esta condição, implica no entanto que durante o processo de aquisição de imagens, a câmara de vídeo permaneça perfeitamente imóvel. Pela experimentação, constata-se que tal não se verifica, sendo frequente o aparecimento de pequenas oscilações que provocam translações de um ou dois pontos da imagem.

Para solucionar este problema propõe-se dotar a condição (4.10) de um mecanismo que inclua a vizinhança do ponto em análise. Deste modo, a máscara primária de movimento (*MPM*) é gerada pela seguinte condição, onde *d* especifica o raio de vizinhança.

$$MPM^{t}(\mathbf{x}) = \begin{cases} 1 & , \text{ se } \arg\max_{R,G,B} \left| \arg\min_{-d \le x,y \le d} \left(I^{t}_{R,G,B}(x,y) - B^{t}_{R,G,B}(x,y) \right) \right| > T \\ 0 & , \text{ caso contrário} \end{cases} \quad (4.11)$$

Concluída a detecção de movimento, especificada por (4.11), é então necessário executar a remoção das regiões, da máscara primária de movimento, afectadas por sombras e brilhos. Este procedimento implica a intercepção da máscara primária de movimento (*MPM*) com a negação das máscaras de sombras e de brilho, como especificado em (4.12).

$$MM^{t} = MPM^{t} \cap \neg MS^{t} \cap \neg ML^{t} \quad (4.12)$$

Porém, tal operação só é realizada após a remoção de ruído, que se manifeste na forma de conjuntos de pontos contíguos de dimensão desprezável, nas máscaras de sombras e de brilhos. Para esse efeito, recorre-se a um operador morfológico de abertura. Esta abertura, composta por uma erosão (Θ), seguida de uma dilatação $(\oplus)$, utiliza dois elementos estruturantes distintos (*A* e *B*, respectivamente).

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (4.13)$$

Assim, as máscaras de sombras e de brilhos, utilizadas no cálculo da máscara de movimento (*MM*) definido por (4.12), são obtidas pelas seguintes operações:

$$MS^{t} = \left(MS^{t} \Theta A\right) \oplus B \quad ML^{t} = \left(ML^{t} \Theta A\right) \oplus B \quad (4.14), (4.15)$$

Na Figura 4.9 apresenta-se um quadro com os resultados obtidos nas várias etapas da segmentação de movimento. Neste exemplo, a imagem de referência (a), mostra um objecto (veículo branco) que não pertence ao plano de fundo. Pela aplicação da condição (4.11) sobre as imagens (a) e (b), obtém-se a máscara primária de movimento (c). Esta máscara encontra-se afectada por sombras e brilhos, que causam deformações na morfologia dos objectos em movimento. As imagens (d) e (e) apresentam respectivamente as máscaras de sombras e brilhos, após a operação morfológica de abertura. Por fim, em (f), apresenta-se a máscara de movimento, livre das perturbações com origem na variação das condições de iluminação.

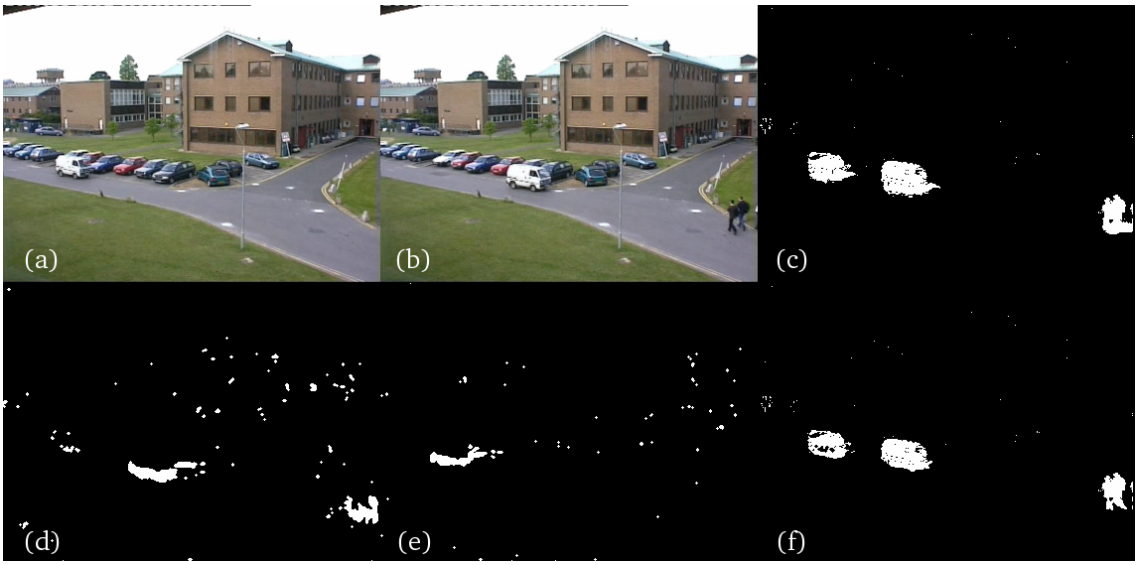


Figura 4.9. (a) Imagem de referência; (b) Imagem alvo; (c) Máscara primária de movimento; (d) Máscara de sombras; (e) Máscara de brilhos; (d) Máscara de Movimento obtida por (4.12).

O último passo refere-se à adaptação do plano de fundo (também designado por imagem de referência). Este processo deve incorporar no plano de fundo as pequenas oscilações verificadas no ambiente (e.g. a lenta variação da intensidade luminosa ao longo do dia), mas não deve todavia assimilar as alterações provocadas pela movimentação de objectos, ou pela ocorrência de sombras e brilhos. Como tal, a presença de movimento, sombras ou brilhos, num determinado ponto da imagem implica uma manutenção do seu valor na coordenada correspondente da imagem de referência. Caso contrário, o valor desse ponto será recalculado por um filtro de resposta a impulso infinito, como especificado na expressão (4.16).

$$B^{t+1}(\mathbf{x}) = \begin{cases} B^t(\mathbf{x}) & , \text{ se } \quad \text{movimento} \vee \text{sombras} \vee \text{brilhos} \\ \alpha \cdot B^t(\mathbf{x}) + (1-\alpha) \cdot I^t(\mathbf{x}) & , \text{ caso contrário} \end{cases} \tag{4.16}$$

4.4. Detecção de Fantasmas e Adaptação do Plano de Fundo

Um problema associado às técnicas de segmentação de movimento baseadas na subtracção por plano de fundo reside na necessidade de se manter uma imagem de referência actualizada, i.e. que assimile as pequenas variações do ambiente. No entanto, essa imagem de referência não deve conter objectos que apresentem movimento, mas apenas objectos estáticos pertencentes ao plano de fundo.

Com a utilização de técnicas de subtracção do plano de fundo é frequente o aparecimento de **fantasmas**. Isto é, falsas detecções provocadas sempre que um objecto, devidamente identificado na imagem de referência como pertencente ao plano de fundo, inicia um movimento de deslocação que o leva a abandonar o espaço anteriormente ocupado. Um exemplo de fantasma pode ser visto na Figura 4.10 (identificado a vermelho).

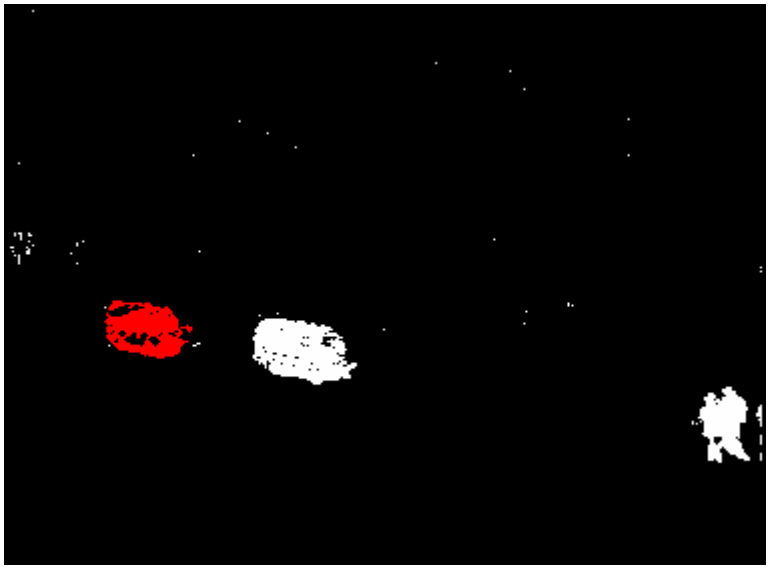


Figura 4.10. Fantasma (a vermelho) resultante do exemplo apresentado na Figura 4.9.

Tipicamente, quando a primeira imagem de referência é adquirida, considera-se que todos os objectos observados correspondem ao plano de fundo. Em ambientes laboratoriais tal condição é facilmente validada. Todavia, em meios reais é impraticável condicionar a presença de objectos, em movimento ou momentaneamente em repouso, que não pertencem ao plano de fundo. Como exemplo veja-se o transtorno que estaria associado à evacuação de todas as pessoas de uma zona de embarque de um aeroporto, de cada vez que fosse necessário reiniciar o sistema de segmentação de movimento responsável pela monitorização daquele local.

Um modo de resolver este problema consiste em detectar a presença de fantasmas nos resultados da segmentação de movimento. Após essa detecção, os pontos da imagem de referência associados a um fantasma, devem então ser actualizados com os valores definidos pela imagem alvo, removendo assim da imagem de referência o objecto que originou o fantasma, como exemplificado na Figura 4.11.

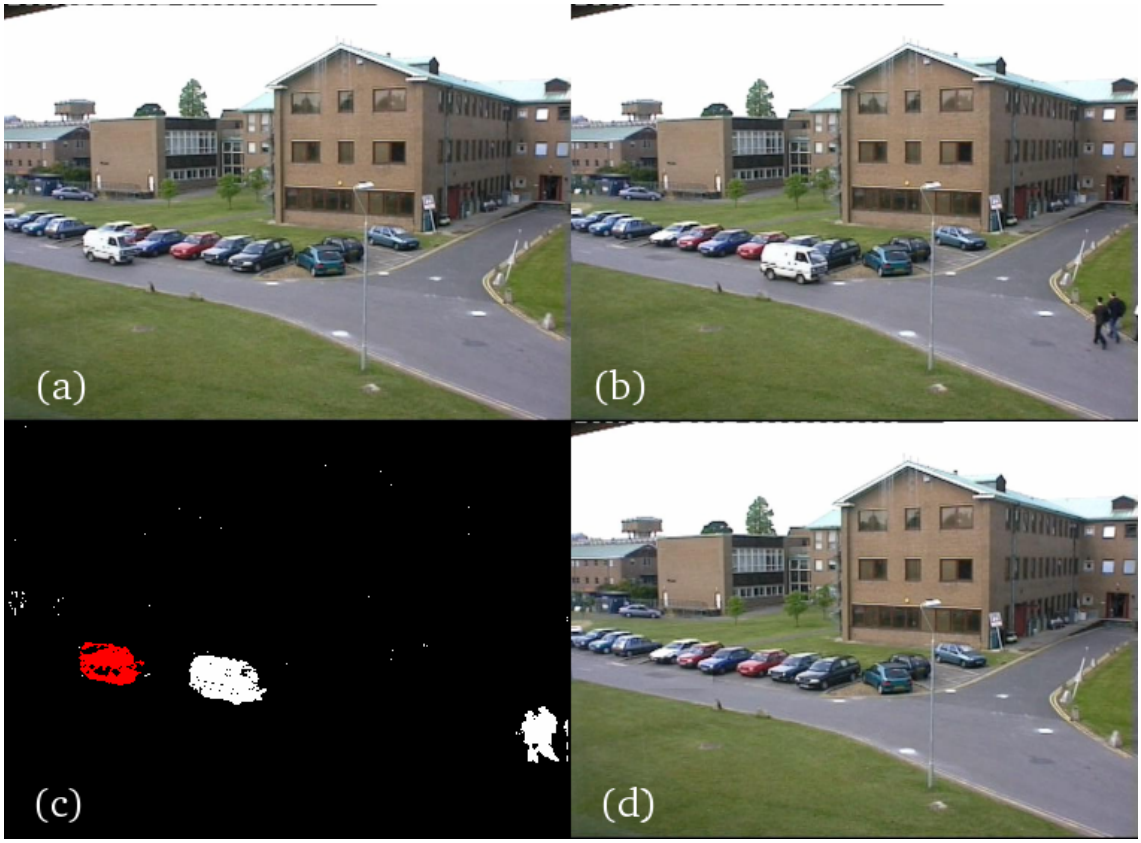


Figura 4.11. (a) Imagem de referência; (b) Imagem alvo; (c) Segmentação de movimento com identificação de fantasma; (d) Imagem de referência actualizada na área afectada por fantasma.

A implementação desta técnica não seria contudo viável se não se garantisse um método de identificação de fantasmas nas regiões segmentadas. Porém, existe uma característica que permite efectuar tal diferenciação. Essa característica é o contorno do objecto em movimento, introduzida por [Kim & Hwang, 2001]. Na abordagem proposta por estes investigadores, recorre-se à informação dos contornos da imagem de referência e dos contornos da imagem alvo, que se fundem posteriormente com os contornos calculados sobre a diferença absoluta entre a imagem alvo e a que lhe precede.

Embora não tenha sido explorada na descoberta de fantasmas, a detecção de contornos de movimento desfruta de uma propriedade com particular relevância na sua identificação. Os contornos de movimento apenas se manifestam nos objectos em deslocação, não sendo

observados nos limites das regiões originadas por fantasmas. Explorando esta propriedade, e combinando a detecção de contornos de movimento com a segmentação de movimento por subtracção de plano de fundo adaptativo, é possível obter um novo método de detecção de fantasmas nos resultados da segmentação.

Não obstante a sua relevância para a detecção dos contornos de objectos em movimento, o método proposto por Kim e Hwang apresenta algumas restrições no processo de inicialização. Nomeadamente, existe a necessidade de eliminar manualmente os contornos de objectos que não pertençam ao plano de fundo, na definição dos contornos da imagem de referência, sendo que esta permanece imutável após a sua inicialização.

Propõe-se assim uma nova técnica para a obtenção dos contornos de movimento, realizada com recurso a três imagens: imagem de referência (B^t); imagem alvo (I^t); e imagem que precede a imagem alvo (I^{t-1}). Esta técnica inicia com o cálculo dos contornos sobre a diferença absoluta entre a imagem alvo e a imagem de referência, para a componente valor (V) do espaço de cor *HSV*. Similarmente, executa-se a identificação dos contornos entre a imagem alvo e a imagem que a precede, como apresentado na Figura 4.12. Os contornos de movimento são então definidos pela intercepção dos resultados obtidos.

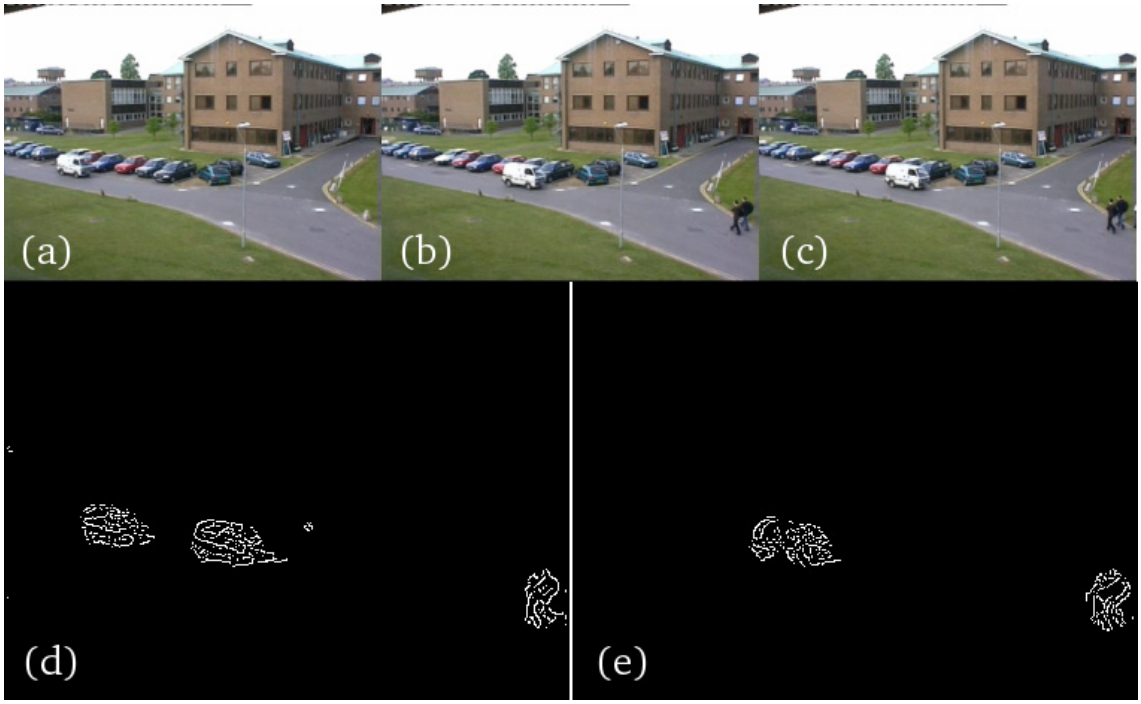


Figura 4.12. (a) Imagem de referência; (b) Imagem que antecede a imagem alvo; (c) Imagem alvo; (d) Contornos de $\left|I_V^t - B_V^t\right|$; (e) Contornos de $\left|I_V^t - I_V^{t-1}\right|$.

4.4.1. Selecção do Detector de Contornos

Para este propósito, consideraram-se os detectores de contornos Prewitt [Prewitt, 1970], Sobel [Sobel, 1978] e Canny [Canny, 1986]. A escolha do algoritmo para a detecção dos contornos é critica no que respeita ao requisito de tempo-real. O seu cálculo deve apresentar resultados aceitáveis, associados a uma reduzida complexidade computacional da operação, i.e. baixo tempo de resposta.

De acordo com os estudos realizados por [Wesolkowski, 1999] e [Shin et al., 2001], verifica-se que estes detectores de contornos apresentam elevada complexidade computacional, apontando o detector Canny como o mais complexo, mas simultaneamente aquele que apresenta melhores resultados. A validade destes pressupostos foi, no âmbito deste trabalho, confirmada pela aplicação dos detectores de contornos sobre o conjunto de imagens de teste, constituída por uma sequência de 2688 imagens (da câmara número 1, do primeiro conjunto de imagens de teste do *PETS* 2001). Para tal, as imagens foram previamente convertidas para a escala de cinzentos (8 *bits* por ponto), com uma resolução de 380 linhas por 280 colunas.

Pela análise dos resultados do teste, verificou-se que os detectores Prewitt e Sobel necessitam de aproximadamente o mesmo tempo para processamento dos contornos (6 milissegundos), enquanto que o detector Canny requer em média 61 milissegundos para executar a mesma função. Sendo necessárias duas operações de contornos para cada nova imagem adquirida, a utilização de um detector Canny na detecção de fantasmas consome nestas condições 122 milissegundos. Este valor ultrapassa em larga escala o período de tempo disponível entre aquisições de imagens no formato *PAL* (40 milissegundos entre imagens).

Com a utilização de detectores Prewitt ou Sobel, o tempo dispendido na detecção de contornos é de aproximadamente 12 milissegundos. Embora seja considerável a diminuição no tempo dispendido para o processamento dos contornos, este processo consome ainda cerca de 30 por cento dos recursos de tempo reservados para todo o processo de segmentação e seguimento de objectos em movimento. Existe assim a necessidade de encontrar um método de detecção de contornos com menor complexidade computacional, mas cujos resultados sejam passíveis de utilização na detecção de fantasmas.

Um detector de contornos simples, eficiente e de reduzido custo computacional, consiste na comparação em cada ponto da componente valor (no espaço de cor *HSV*) com quatro pontos vizinhos. Se a diferença absoluta entre o ponto e um dos seus vizinhos for superior a um determinado valor, então esse ponto é marcado como contorno. Baseada nesta descrição, a seguinte função irá produzir uma máscara binária com os contornos observados:

$$\Phi(I(x,y)) = \begin{cases} 1 \quad , \text{ se } \quad |I(x,y) - I(x-1,y-1)| > T \quad \wedge \\ \qquad\qquad |I(x,y) - I(x-1,y+1)| > T \quad \wedge \\ \qquad\qquad |I(x,y) - I(x+1,y-1)| > T \quad \wedge \\ \qquad\qquad |I(x,y) - I(x+1,y+1)| > T \\ 0 \quad , \text{ caso contrário} \end{cases} \tag{4.17}$$

De modo a avaliar o desempenho do detector de contornos definido pela expressão (4.17) no que se refere ao tempo de execução, este foi submetido a um teste para processamento do mesmo conjunto de imagens utilizado na avaliação dos detectores Prewitt, Sobel e Canny. Pela medição do período de tempo necessário ao processamento de contornos de cada imagem, verificou-se que o detector de contornos proposto necessita em média de 2 milissegundos por imagem. O recurso a este detector traduz-se assim numa optimização dos recursos de tempo, na ordem dos 67 por cento em relação aos detectores Prewitt e Sobel. O detector identificado pela função (4.17) é assim seleccionado como o mais adequado ao processamento da detecção de contornos, neste caso particular.

4.4.2. Contornos de Movimento e Identificação de Fantasmas

Como referido anteriormente, a detecção dos contornos em movimento obtém-se pela aplicação da função (4.17) ao resultado da diferença, para as componentes de valor do espaço de cor *HSV*, entre a imagem de referência e a imagem alvo, e ao resultado da diferença entre a imagem alvo e a imagem que a precede. A intersecção das duas máscaras binárias resultantes destas operações encerra assim os contornos dos objectos em movimento. Este processo pode ser descrito pela seguinte expressão:

$$MCM^t = \Phi\left(\left|I_V^t - B_V^t\right|\right) \cap \Phi\left(\left|I_V^t - I_V^{t-1}\right|\right) \tag{4.18}$$

Após o cálculo da máscara de contornos de movimento (*MCM*), é necessário realizar a contagem e etiquetagem das regiões de pontos segmentados e identificados como tal na máscara de movimento (*MM*). Para cada região, prossegue-se com o cálculo do seu centro-de-massa e área. As regiões cujas áreas sejam inferiores a um mínimo pré-definido são consideradas como ruído sendo removidas da máscara de movimento.

Concluída a etiquetagem das regiões segmentadas pela máscara de movimento, prossegue-se com a computação dos seus contornos. Dado se tratar de uma máscara binária (identificando movimento ou plano de fundo), a detecção desses contornos é executada por um algoritmo semelhante ao proposto por [Freeman, 1970], de reduzida carga computacional. A imagem resultante, denominada por máscara de contornos da segmentação (*MCS*), inclui os contornos dos objectos em movimento e os contornos dos fantasmas.

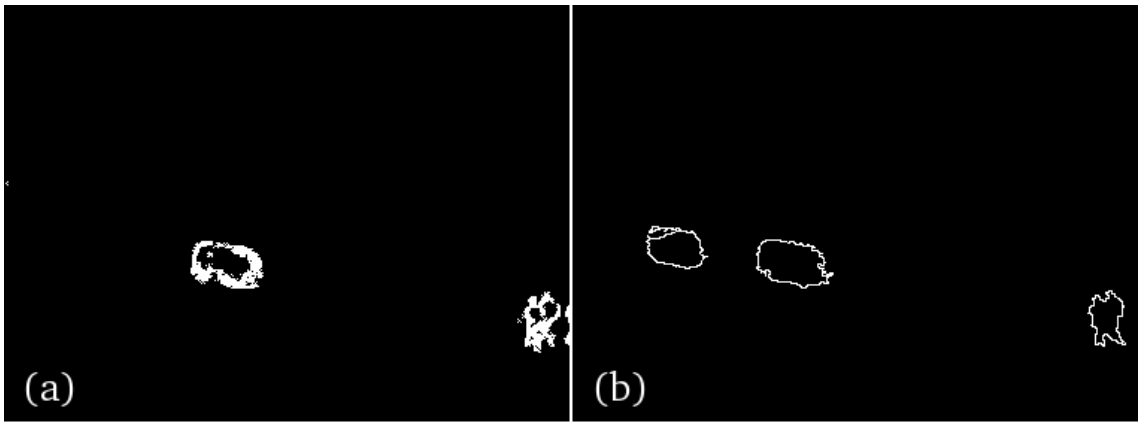


Figura 4.13. (a) Exemplo de uma máscara de contornos de movimento (*MCM*); (b) Máscara de contornos da segmentação (*MCS*).

A detecção de fantasmas é então realizada, para cada região segmentada, através do cálculo da razão entre o perímetro obtido pela intercepção da *MCM* com a *MCS*, e o perímetro da região. Se a relação entre estes valores for inferior a 10%, a região é assinalada como um fantasma:

$$região(i) = \begin{cases} fantasma & , \text{ se } \dfrac{Perímetro\left(MCM^t(i) \cap MCS^t(i)\right)}{Perímetro\left(MCS^t(i)\right)} < 10\% \\ movimento & , \text{ caso contrário} \end{cases} \quad (4.19)$$

4.4.3. Adaptação do Plano de Fundo à Detecção de Fantasmas

Por fim, a imagens de referência são actualizadas. Isto porque são necessárias duas imagens de referência, uma no espaço de cor *RGB* e outra no espaço *HSV*. Esta actualização é efectuada de modo a que os pontos identificados como sombras, brilhos ou movimento, mantenham inalteradas as suas componentes de cor nas imagens de referência.

Os pontos contidos em regiões classificadas como fantasmas são actualizados, nas imagens de referência, com os valores das componentes de cor definidas para esses pontos na imagem adquirida. Os restantes pontos são actualizados com um filtro de resposta a impulso infinito, de forma a adaptar a imagem de referência a pequenas variações de luminosidade.

A adaptação do plano de fundo pode, por conseguinte, ser expressa por:

$$B^{t+1}(\mathbf{x}) = \begin{cases} B^t(\mathbf{x}) & , \text{ se } \quad \text{movimento} \vee \text{sombras} \vee \text{brilhos} \\ I^t(\mathbf{x}) & , \text{ se } \quad \text{fantasma} \\ \alpha \cdot B^t(\mathbf{x}) + (1-\alpha) \cdot I^t(\mathbf{x}) & , \text{ caso contrário} \end{cases} \tag{4.20}$$

Onde, α controla a celeridade na adaptação a alterações de iluminação.

4.5. Avaliação da Técnica de Segmentação Proposta

Na avaliação da técnica de segmentação aqui proposta, existem dois factores cuja análise é relevante: tempo de execução e qualidade da segmentação. Enquanto que o primeiro factor é facilmente obtido pela medição do tempo médio requerido pelo algoritmo na segmentação de movimento numa sequência de imagens, a análise da qualidade da segmentação acarreta maior complexidade.

De acordo com o que foi referido na Secção 2.1.1, a avaliação quantitativa de uma técnica de segmentação de movimento, exige a existência de um conjunto de dados de referência que identifique os pontos que definem, em cada imagem, a forma dos objectos que se encontram em movimento. Este modelo de avaliação permite a quantificação dos: **falsos positivos** (FP); **falsos negativos** (FN), **verdadeiros negativos** (VN); e **verdadeiros positivos** (VP). Sendo que os falsos positivos e os falsos negativos referem-se respectivamente aos pontos classificados erradamente como movimento e plano de fundo.

Os verdadeiros negativos identificam as correctas classificações de plano de fundo, enquanto que os verdadeiros positivos quantificam as classificações de pontos de movimento realizadas com sucesso.

Apesar da divulgação efectuada pelos promotores da *PETS*, sobre a existência de um conjunto de dados de referência para a segmentação de movimento [PETS, 2006], contendo 11.2% das imagens da câmara número 1, este conjunto de dados não foi no entanto disponibilizado. Como tal, e dado que a construção de um conjunto de dados de referência não se encontrava definido no plano de trabalhos deste doutoramento, não será realizada uma avaliação quantitativa da técnica de segmentação de movimento. Contudo, a avaliação da técnica de seguimento, através dos dados de referência fornecidos pela *PETS*, possibilitará uma avaliação global do sistema, i.e., técnica de segmentação e seguimento de objectos em movimento.

Embora não seja possível realizar uma avaliação quantitativa do processo de segmentação de movimento, a demonstração da qualidade da técnica proposta pode ser observada através de imagens que definem os resultados da segmentação para o conjunto de imagens de teste. Na Figura 4.14 e na Figura 4.15 apresentam-se os resultados obtidos com a técnica de segmentação proposta, para algumas imagens da sequência de teste da *PETS* 2001.

As regiões segmentadas (áreas sinalizadas a branco) correspondem efectivamente aos objectos em movimento. Verifica-se ainda que fenómenos como sombras e brilhos são eliminados com sucesso do resultado da segmentação.

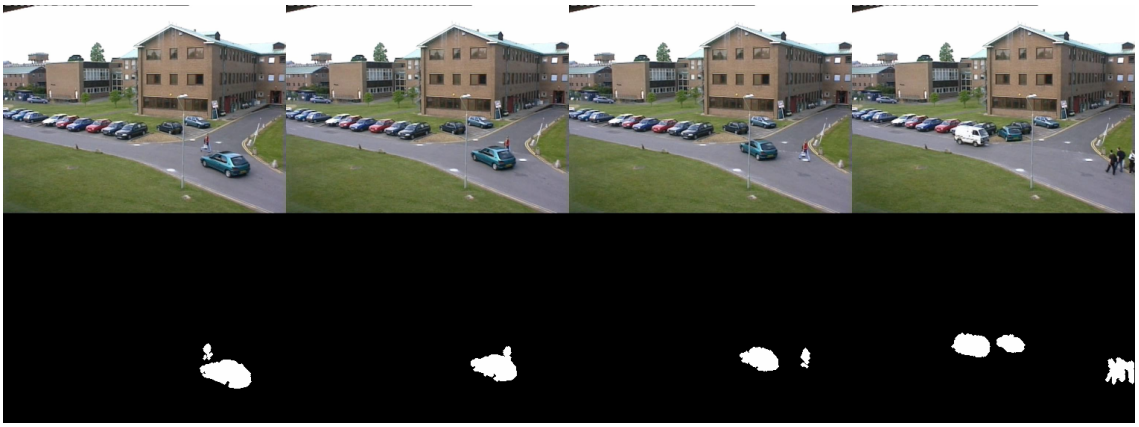


Figura 4.14. Resultados da segmentação de movimento para as imagens número 551, 575, 598 e 800 da sequência de teste.

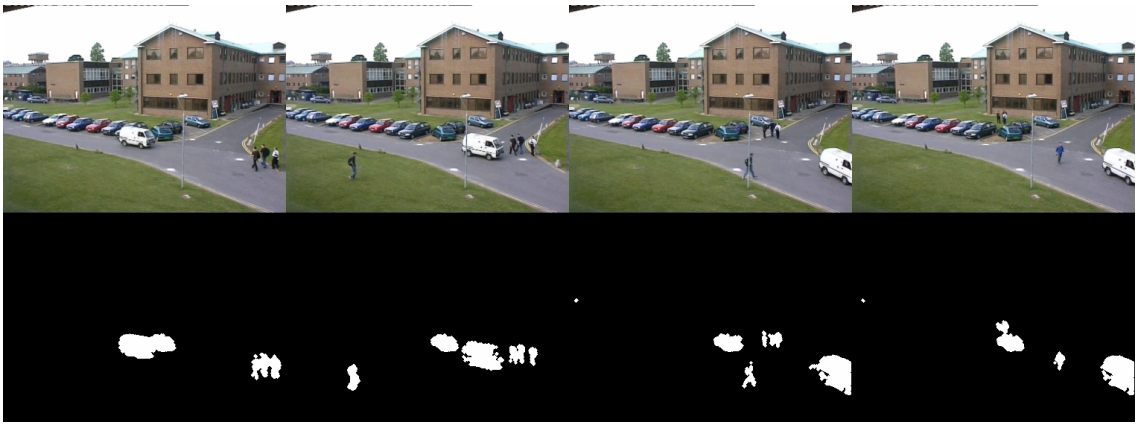


Figura 4.15. Resultados da segmentação de movimento para as imagens número 814, 880, 989 e 1160 da sequência de teste.

Para avaliar o tempo de execução da técnica de segmentação, bem como das componentes que a constituem, recorreu-se à sequência de imagens da câmara número 1, do primeiro conjunto de imagens de teste da *PETS* 2001. Assim, as 2688 imagens com uma resolução de 380 por 280 pontos foram processadas num computador com processador a 3.0GHz, de 32 *bits*, com uma frequência de barramento a 800MHz, equipado com 512MB de memória RAM.

Realizaram-se seis medições do tempo de processamento da técnica de segmentação para cada imagem da sequência de teste. O tempo de execução médio foi então calculado, resultando na curva assinalada a vermelho na Figura 4.16. Como se pode verificar, durante a inicialização do sistema (primeira imagem adquirida) o tempo dispendido é consideravelmente superior ao necessário no processamento das restantes imagens da sequência. Tal facto deve-se à alocação de memória executada durante a inicialização do sistema, bem como à geração do plano de fundo. Por este motivo, o tempo requerido no processamento da primeira imagem não deve ser considerado para o cálculo do tempo médio da técnica de segmentação.

Com base nas medições efectuadas, obteve-se um valor de tempo médio de execução da técnica de segmentação de 33190μs por imagem, com um desvio padrão de 226μs. De modo a avaliar uma possível relação entre a tempo de segmentação e a área de pontos segmentados, realizou-se a medição desta área para cada imagem da sequência (curva a azul na Figura 4.16). Pela análise do coeficiente de correlação de *Pearson* entre o tempo de execução e a área segmentada, verificou-se que na técnica proposta existe uma baixa

correlação (0.42) entre estas duas variáveis. Esta característica da técnica apresentada garante ao sistema estabilidade quando defrontado com ambientes de extremo movimento.

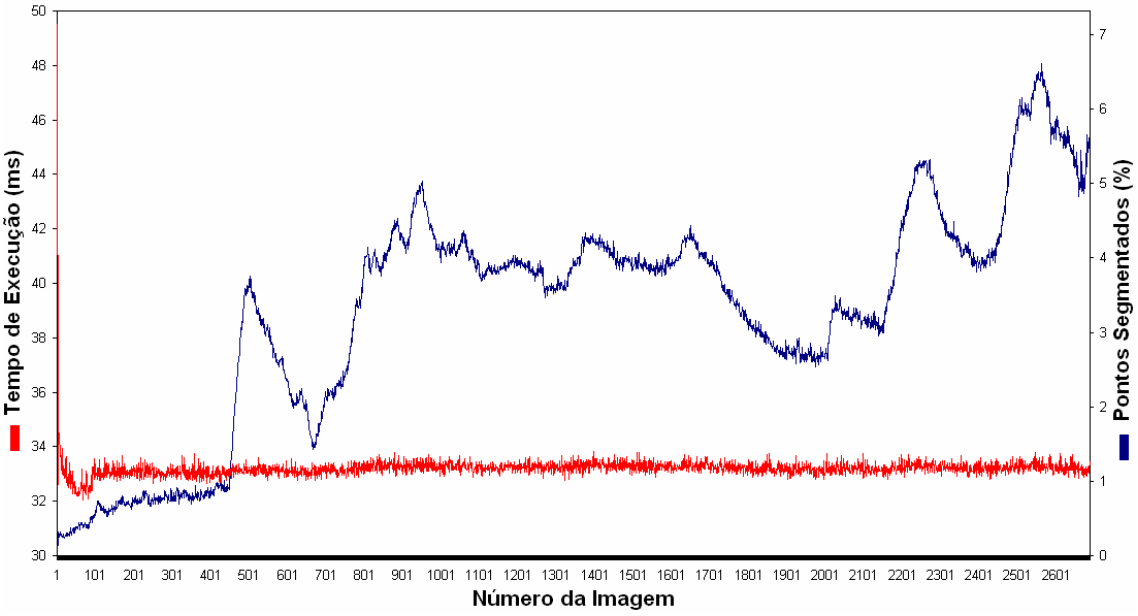


Figura 4.16. Tempo de execução (a vermelho) e percentagem de pontos segmentados (a azul) por imagem da sequência.

4.6. Discussão

Neste capítulo, propôs-se uma nova técnica tempo-real para a segmentação de objectos em movimento. A abordagem apresentada mostrou ser passível de aplicação em ambientes reais, onde a interferência das condições de iluminação, i.e. sombras e brilhos, constituem frequentemente factores de enfraquecimento da precisão dos resultados. Um outro princípio de especial relevância, aqui abordado, tem em conta os problemas associados à inicialização dos sistemas de segmentação por subtracção de plano de fundo, ou seja, a detecção e a remoção de fantasmas, bem como a utilização dessa informação na correcta adaptação do plano de fundo.

A detecção de sombras e brilhos, realizada a partir do espaço de cor *HSV*, mostrou ser eficaz em aplicações de vídeo-vigilância onde as imagens são monitorizadas por uma câmara de vídeo fixa. Contudo, não é possível empregar este método no processamento de imagens provenientes de câmaras rotativas ou que permitam efectuar variações na ampliação da cena observada. Isto porque a abordagem proposta requer a existência de um

plano de fundo estático, de modo a permitir efectuar a análise da variação cromática em cada ponto da imagem.

Apesar de possuir quatro parâmetros de funcionamento (variação da saturação (τ_S), variação da matriz (τ_H), α, e β), a detecção de sombras e brilhos apenas necessita que se realizem pequenos ajustes em dois destes parâmetros (α e β). Embora não tenha sido proposta nenhuma técnica para a definição dos valores dos parâmetros α e β que optimizam o desempenho do método de detecção de sombras e brilhos, foram no entanto anunciados intervalos de valores para os quais se observam melhores resultados.

A proposta de um método de detecção de fantasmas para técnicas de segmentação baseadas na subtracção do plano de fundo permite minorar o tempo e esforço exigidos durante a inicialização do sistema. Assim, com a introdução deste método, o estado em que os objectos se encontram no arranque do sistema (em movimento ou repouso) deixa de ter relevância. A identificação manual dos objectos que pertencem ao plano de fundo deixa igualmente de ser necessária, permitindo assim automatizar em pleno o processo de segmentação de objectos em movimento.

Capítulo 5

5. Seguimento de Objectos em Movimento

Após a segmentação de objectos em movimento, é necessário efectuar o seu seguimento durante o período que estes se mantenham em cena. Para este efeito, foi escolhida uma técnica de seguimento baseada em *Modelos de Aparência*. Esta técnica permite a realização do seguimento de objectos deformáveis, mostrando-se eficaz em situações de oclusão, fusão e separação de objectos.

O presente capítulo abre com a definição dos *Modelos de Aparência*. A técnica proposta para o seguimento de objectos em movimento é apresentada principiando pela construção de uma matriz de correspondência entre as regiões segmentadas e os trilhos existentes. Avança-se com o modelo empregue no alinhamento dos trilhos às regiões segmentadas, seguindo-se a apresentação do método de atribuição de pontos aos trilhos concorrentes. Por fim, descreve-se a abordagem adoptada na actualização dos *Modelos de Aparência*, terminando com a apresentação da técnica proposta para o seguimento de objectos resultantes da desagregação de elementos de um trilho.

O capítulo prossegue com a avaliação da técnica proposta, tendo por base os dados de referência para o seguimento de objectos disponibilizados pela *PETS* 2001. Encerra-se com uma discussão onde se enunciam as principais vantagens e inconvenientes do sistema proposto.

5.1. Modelos de Aparência

Um modo de realizar o seguimento dos objectos ao longo do tempo, segmentados a partir de uma sequência de imagens digitalizadas, consiste na utilização de *Modelos de Aparência*. Inicialmente introduzidos por [Haritaoglu et al., 2000] no processamento de sequências de imagens em escala de cinzentos, adquiridas por câmaras monoculares fixas, os *Modelos de Aparência* podem ser entendidos como representações dinâmicas de memória das características visuais (cor e forma) que identificam os objectos monitorizados.

A utilização desta técnica requer que, durante a operação de seguimento de um objecto, exista um *Modelo de Aparência* que o caracterize. Este modelo deve ser mantido enquanto o objecto se mantiver em cena, e actualizado a cada nova imagem adquirida de forma a incorporar novas informações que possibilitem discriminar o objecto.

Cada *Modelo de Aparência* é composto por dois tipos de informação: *Imagem de Aparência* (*IA*); e *Máscara de Probabilidade* (*MP*). A *Imagem de Aparência* mantém um modelo de cor do objecto observado, enquanto que a *Máscara de Probabilidade* representa a forma mais provável que o objecto pode assumir. Um exemplo de um *Modelo de Aparência* pode ser visto na Figura 5.1.

Apesar da complexidade computacional que a técnica impõe, a ideia que está na sua génese é bastante elementar. O processo de segmentação de movimento fornece um conjunto de regiões que identificam a localização e a forma dos objectos que se encontram em movimento. Através do cruzamento de informação das regiões segmentadas (máscara de movimento) com a imagem adquirida, é possível obter um conjunto de objectos identificados pela sua forma e cor. Pela análise da similaridade entre um *Modelo de Aparência* de um determinado objecto, com o conjunto constituído pelas regiões segmentadas, é então possível efectuar uma correspondência entre cada região e um *Modelo de Aparência* de um objecto.

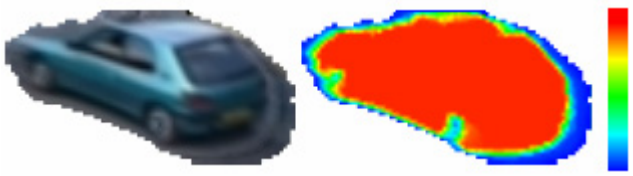


Figura 5.1. Exemplo de um *Modelo de Aparência*. À esquerda a *Imagem de Aparência*, e à direita a *Máscara de Probabilidade* acompanhada pela respectiva escala.

5.1.1. Matriz de Correspondência

Do processo de segmentação de movimento resulta um conjunto de regiões segmentadas. Cada uma destas regiões de pontos da imagem define a área bidimensional que descreve a forma de um objecto em movimento. Como resultado, a cada nova imagem processada pela técnica de segmentação, obtém-se uma lista de regiões:

$$R = \{r_1, r_2, r_3, r_4, ..., r_J\} \tag{5.1}$$

Cada região apresentada na lista é então caracterizada por uma máscara binária (M_j) que identifica os pontos que a constituem, e por um rectângulo (BB_j) que envolve todos os pontos da referida máscara. Assim, uma região r_j é definida por:

$$r_j = \{M_j, BB_j\} \tag{5.2}$$

De modo a efectuar o seguimento de um objecto, desde a sua entrada em cena até ao abandono da área sob vigilância, é necessário gerar e manter um trilho próprio a esse objecto. Cada trilho é então caracterizado por um identificador que o distingue dos trajectos dos restantes objectos observados. Para além dessa informação, é ainda necessário que se agregue dados sobre o tipo de objecto, a posição e área que este ocupou na última observação, bem como o *Modelo de Aparência* que o distingue. Deste modo, a cada nova imagem adquirida deve existir um conjunto de trilhos, tal que:

$$T = \{t_1, t_2, t_3, t_4, ..., t_K\} \tag{5.3}$$

com,

$$t_k = \{id, tipo_k, BB_k, IA_k, MP_k\} \tag{5.4}$$

onde *id* é o identificador do trilho, *tipo* especifica a classe do objecto (i.e. "pessoa", "grupo" ou "veículo"), BB_k corresponde à posição e área do rectângulo que encerra o objecto na sua última observação, IA_k e MP_k especificam respectivamente a *Imagem de Aparência* e a *Máscara de Probabilidade* do objecto.

Com esta abordagem, para cada instante de tempo existem dois conjuntos. Um conjunto de J elementos que contém as regiões segmentadas e um conjunto de trilhos com K objectos. Deste modo, é possível construir uma matriz de correspondência C que associe a

cada trilho t_k uma região r_j. Esta matriz, com dimensão *J*x*K*, onde *J* define o número de regiões segmentadas e *K* o número de trilhos existentes, é uma matriz binária que atribui o valor unitário ao elemento $C_{j,k}$ sempre que a região *j* esteja associada ao trilho *k*.

A associação entre regiões e trilhos é estabelecida sempre que se verifique a sobreposição da região segmentada de um objecto (BB_j) pela área definida na última observação de um trilho (BB_k). A construção da matriz de correspondência com dimensão *J*x*K* é então expressa pela seguinte condição:

$$C_{j\in J,k\in K} = \begin{cases} 1 & , \text{ se } \quad BB_j \cap BB_k \neq \varnothing \\ 0 & , \text{ caso contrário} \end{cases} \tag{5.5}$$

Nos casos em que uma região não é associada a qualquer trilho, um novo trilho deve então ser gerado e atribuído a esse objecto. Tais situações podem dever-se à ausência de uma lista de trilhos, ou seja, à inexistência qualquer objecto em movimento nas imagens predecessoras. A formação de um novo trilho pode também aplicar-se aquando do surgimento de um novo objecto em cena.

Sempre que um novo trilho é gerado, é-lhe conferido um identificador numérico que o permitirá diferenciar dos restantes trajectos observados. De modo a evitar problemas de *overflow* na atribuição do identificador a cada novo trilho, a aplicação deve manter em memória uma lista de identificadores disponíveis. Esta lista deve ser actualizada sempre que um trilho seja gerado ou removido.

Para além do identificador, os trilhos originados pelo surgimento de novos objectos devem iniciar a *Imagem de Aparência* com o conjunto de pontos em formato *RGB* do objecto segmentado, sendo que a *Máscara de Probabilidade* deve principiar com um valor intermédio contido no intervalo de valores admitidos [0, 1], nomeadamente:

$$\begin{aligned} IA_k(\mathbf{x}) &= I_{R,G,B}(\mathbf{x}) \quad , \forall \mathbf{x} \in M_j \\ MP_k(\mathbf{x}) &= \begin{cases} 0.5 & , \text{ se } \quad \mathbf{x} \in M_j \\ 0 & , \text{ caso contrário} \end{cases} \end{aligned} \tag{5.6}$$

A atribuição da classe a que o objecto pertence (i.e. "pessoa", "grupo" ou "veículo") não se realiza no momento da criação de um novo trilho. De modo a evitar erros na classificação do tipo do objecto, originados pela fraca representação do objecto nos instantes iniciais, a identificação da classe só se realiza após um determinado número de imagens.

Como vimos, o processo de segmentação de movimento disponibiliza uma lista de regiões que identificam os objectos em movimento. Embora seja frequente a representação de um objecto por uma única região, em determinados casos uma região segmentada pode conter dois ou mais objectos (e.g. um grupo de pessoas contém vários indivíduos). Tais situações têm origem na oclusão parcial entre vários objectos em movimento na área observada. Por outro lado, um único objecto pode também ser identificado por várias regiões segmentadas. Neste último caso, ao conjunto das regiões que definem um objecto dá-se o nome de **macro-objecto** (*MO*). Pela matriz de correspondência *C*, um macro-objecto é definido por uma região ou pelo conjunto regiões segmentadas associados a um trilho comum.

De forma a clarificar os conceitos aqui apresentados, considere-se o seguinte exemplo, em que do processo de segmentação resulta um conjunto composto por quatro regiões, que terão de ser avaliadas segundo um grupo de quatro trilhos previamente identificados. De tal modo que:

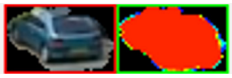
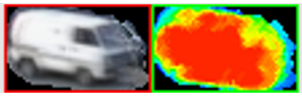
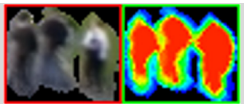
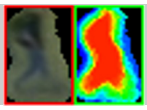
$$R = \{r_1, r_2, r_3, r_4\}, \quad T = \{t_1, t_2, t_3, t_4\} \tag{5.7}$$

Considere-se ainda que as regiões (*R*) identificadas no processo de segmentação de movimento, e os trilhos existentes (*T*), são caracterizados respectivamente pelas informações contidas na Tabela 5.1, e na Tabela 5.2.

Tabela 5.1. Exemplo de conjunto de regiões segmentadas para a imagem número 865 do conjunto de dados de treinos da *PETS* 2001.

R	***BB***		**Máscara Binária**
	(x_0,y_0)	**Dimensão**	
r_1	(194,161)	80x46	
r_2	(306,178)	25x30	
r_3	(332,180)	15x29	
r_4	(57,201)	19x37	

Tabela 5.2. Exemplo de lista de trilhos, após o processamento de 865 das imagens de treino da *PETS* 2001.

T	**ID**	**Tipo**	***BB***		**Modelo de Aparência**	
			(x_0,y_0)	**Dimensão**	***IA***	***MP***
t_1	00	Veículo	(194,161)	40x26		
t_2	01	Veículo	(219,172)	55x35		
t_3	02	Grupo	(303,174)	44x37		
t_4	03	Pessoa	(55,201)	27x38		

De acordo com a condição (5.5), obtém-se uma matriz de correspondência *C*, como representada em (5.8), onde as linhas se referem às regiões segmentadas e as colunas identificam os trilhos. Pela análise dos valores exibidos pela matriz de correspondência, verifica-se que a primeira região segmentada (r_1) consiste numa fusão de dois trilhos (t_1 e t_2). Por outro lado, as regiões r_2 e r_3 (segunda e terceira linha da matriz) derivam de uma separação do trilho número t_3 (terceira coluna). A quarta região (r_4) é atribuída exclusivamente ao trilho t_4.

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.8}$$

Após a geração da matriz de correspondência, o passo seguinte consiste na elaboração do conjunto dos macro-objectos. De acordo com o referido anteriormente, um macro-objecto é formado por uma região ou conjunto de regiões associadas a um mesmo trilho. Tomando como exemplo a matriz de correspondência apresentada em (5.8), obtém-se um conjunto de macro-objectos que será posteriormente utilizado no cálculo do alinhamento dos trilhos. Assim, da matriz (5.8) resultam três macro-objectos:

$$MO = \{mo_1, mo_2, mo_3\} \tag{5.9}$$

onde,

$$mo_1 = \{\{r_1\},\{t_1,t_2\}\}, \quad mo_2 = \{\{r_2,r_3\},\{t_3\}\}, \quad mo_3 = \{\{r_4\},\{t_4\}\} \tag{5.10}$$

5.1.2. Alinhamento dos Trilhos

Após a correcta associação entre os trilhos e as regiões segmentadas, prossegue-se com o alinhamento dos trilhos, de acordo com o conjunto de macro-objectos gerado no passo anterior. O alinhamento é efectuado através do cálculo, para cada trilho, do deslocamento que melhor o ajusta a um determinado macro-objecto. Este processo é executado para todos os trilhos associados a um macro-objecto, principiando por aqueles que se encontram mais próximos da câmara de vídeo.

O cálculo do deslocamento é obtido pela maximização da função de ajuste P_{FIT}:

$$P_{FIT}(mo_z, t_k, \delta) = \frac{\sum_{x,y \in mo_z} P_{APP}\left(I_{R,G,B}(x,y), IA_k(x-\delta x, y-\delta y)\right) \cdot MP_k(x-\delta x, y-\delta y)}{\sum_{x,y \in t_k} MP_k(x,y)} \tag{5.11}$$

$$\delta = (\delta x, \delta y) = \arg\max_{\delta}\left(P_{FIT}(mo_z, t_k, \delta)\right) \tag{5.12}$$

Considerando que as componentes de cor do espaço *RGB* não são correlacionadas, e possuem igual variância, então:

$$P_{APP}(I_{R,G,B}, \mu_{R,G,B}) = \frac{1}{\sqrt{(2\pi)^3 \cdot \sigma^3}} \cdot e^{-\frac{1}{2}\cdot\left[\left(\frac{I_R-\mu_R}{\sigma}\right)^2+\left(\frac{I_G-\mu_G}{\sigma}\right)^2+\left(\frac{I_B-\mu_B}{\sigma}\right)^2\right]} \tag{5.13}$$

Inicialmente testa-se a função de ajuste (5.12) para um deslocamento nulo (i.e. δx=0, δy=0). Posteriormente, são utilizados dezasseis vectores de deslocamento (5.14) distintos, cobrindo uma área de 9x9 pontos centrada na posição do ponto que reflecte a localização do objecto na imagem anterior.

$$\begin{aligned}\vec{\delta} = \{&(-4,-4),(0,-4),(4,-4),(-2,-2),(0,-2),(2,-2),(-4,0),\\ &(-2,0),(2,0),(4,0),(-2,2),(0,2),(2,2),(-4,4),(0,4),(4,4)\}\end{aligned} \tag{5.14}$$

Concluído o cálculo da função de ajuste para os dezassete deslocamentos propostos, realiza-se então uma translação da região de interesse de acordo com o valor do vector de deslocamento que maximizou a função P_{FIT}. O processo de cálculo da função de ajuste para os vectores de deslocamento especificados em (5.14), bem como a translação da região de interesse, voltam-se a repetir enquanto se verificar um aumento da função de ajuste.

Depois de obtido o vector de deslocamento que maximiza a função de ajuste, os pontos do macro-objecto que coincidem com o trilho em análise são excluídos dos processos de ajuste dos restantes trilhos candidatos.

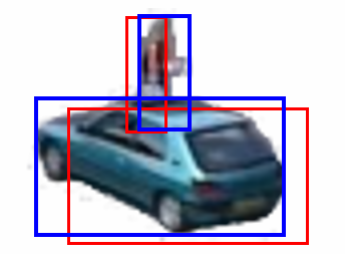


Figura 5.2. Representação de alinhamento de dois trilhos sobre uma única região segmentada composta por um veículo e uma pessoa. Os rectângulos a vermelho indicam a posição dos objectos na imagem anterior. A azul representam-se as novas posições, depois de realizado o alinhamento.

5.1.3. Atribuição de Pontos a Trilhos

Um macro-objecto pode caracterizar um único objecto físico (e.g. uma pessoa) ou um conjunto de objectos, como exemplificado na Figura 5.2. Neste último caso é necessário realizar a atribuição dos pontos de um macro-objecto aos trilhos que lhe estão associados. Esta atribuição é executada de modo a que cada ponto pertencente a um macro-objecto, com um valor não nulo na máscara de probabilidade, seja atribuído a um dos trilhos que competem por esse ponto.

A associação de cada ponto a um trilho é efectuada através do cálculo do produto da probabilidade do ponto pertencer à *Imagem de Aparência* do trilho, pela probabilidade de pertencer ao trilho. Assim, o ponto com coordenadas (x,y), pertencente a uma máscara binária associada ao macro-objecto (mo_z) em análise, será atribuído ao trilho $t_k \in mo_z$ que maximize a função:

$$\arg\max_k \left(P_{APP}\left(I_{R,G,B}(x,y), AI_k(x-\delta x, y-\delta y) \right) \cdot PM_k(x-\delta x, y-\delta y) \right) \qquad (5.15)$$

onde $(\delta x,\delta y)$ especifica o vector de deslocamento obtido após o alinhamento dos trilhos, de acordo com o processo descrito na Secção 5.1.2.

Figura 5.3. Atribuição de pontos de um macro-objecto para dois trilhos concorrentes.

Através deste processo, cada ponto de um macro-objecto é atribuído a uma de três classes possíveis. Se o ponto estiver associado a um **trilho** t_k, então diz-se que este pertence ao conjunto de pontos atribuídos (PA_k) a esse trilho. Se por outro lado, o ponto identificado no macro-objecto não possui um valor positivo na *Máscara de Probabilidade*, é então qualificado como **novo**. Caso contrário, o ponto é classificado como **perdido**.

Através do conhecimento da área composta pelo conjunto de pontos atribuídos (PA_k), é possível identificar o termo de um trilho. Se a razão, entre o número de pontos atribuídos pelo número de pontos da *Máscara de Probabilidade* com valor diferente de zero for inferior a um valor predefinido, o trilho é assinalado como perdido. Ou seja,

$$\frac{\sum_{(x,y)\in t_k} PA_k(x,y)}{\sum_{(x,y)\in t_k} MP_k(x,y)} < 0.06 \tag{5.16}$$

A perda de um trilho não implica que este seja eliminado do processo de seguimento de objectos. A falha na detecção de trilhos pode ter origem na oclusão total do objecto observado, quer seja por outro objecto em movimento, ou por um qualquer obstáculo pertencente à cena. Deste modo, um trilho só deve ser eliminado do sistema após uma sequência de falhas sucessivas.

Quando um trilho é removido da memória do sistema de seguimento, o *Modelo de Aparência* que lhe estava associado é também eliminado. Como efeito, se o objecto reentrar em cena, este será detectado como uma nova entidade e não como uma extensão do movimento previamente observado. Para contornar este problema, pode-se utilizar uma abordagem que mantém os *Modelos de Aparência* de trilhos perdidos, durante um período de tempo predefinido. Contudo, a memoria RAM torna-se num recurso crítico quando se trabalha com um número elevado de objectos, possibilitando apenas a manutenção desta informação durante curtos espaços de tempo.

5.1.4. Actualização dos Modelos de Aparência

O processo de seguimento de objectos continua com a actualização dos *Modelos de Aparência* para os trilhos observados. Nesta fase, são actualizadas a *Imagem de Aparência* (IA_k), a *Máscara de Probabilidade* (MP_k), a posição e área (BB_k) de cada trilho isolado. Para trilhos que tenham sofrido fusão, apenas se realiza a actualização da posição do objecto. Os *Modelos de Aparência* são actualizados de acordo com as seguintes regras:

$$IA_k^{t+1}(x,y) = \begin{cases} \alpha \cdot IA_k^t(x,y) + (1-\alpha) \cdot I(x,y) & \text{, se } (x,y) \in PA_k \\ I(x,y) & \text{, se novo} \\ IA_k^t(x,y) & \text{, caso contrário} \end{cases} \tag{5.17}$$

$$MP_k^{t+1}(x,y) = \begin{cases} \alpha \cdot MP_k^t(x,y) + (1-\alpha) & \text{, se } (x,y) \in PA_k \\ 0.5 & \text{, se novo} \\ \alpha \cdot MP_k^t(x,y) & \text{, caso contrário} \end{cases} \tag{5.18}$$

onde $0 \leq \alpha \leq 1$. A posição e área são actualizadas conforme especificado em (5.19) e (5.20).

$$(x_0, y_0) = (x_0 + \delta x, y_0 + \delta y) \tag{5.19}$$

$$\begin{aligned} \text{largura} &= \max_{MP_k(x,y)\neq 0}(x) - \min_{MP_k(x,y)\neq 0}(x) \\ \text{altura} &= \max_{MP_k(x,y)\neq 0}(y) - \min_{MP_k(x,y)\neq 0}(y) \end{aligned} \tag{5.20}$$

5.1.5. Desagregação de Elementos de um Trilho

Os objectos observados sofrem frequentes alterações durante o período em que se encontram em cena. As alterações de forma e cor de um objecto são, de acordo com o especificado no passo anterior, incorporadas no *Modelo de Aparência* do trilho que lhe está associado. Todavia, existe um outro fenómeno que não foi ainda considerado. Designadamente, a separação de objectos que partilham um mesmo trilho.

Quando as pessoas se deslocam em grupos, encontram-se habitualmente de tal forma próximas umas das outras que a segmentação é efectuada por uma única região. Existem no entanto outros casos em que um objecto em movimento pode ser decomposto em vários elementos, e.g. a um veículo em movimento está sempre associado uma ou mais pessoas (condutor e passageiros), que podem a qualquer momento abandonar o veículo.

Nos dois casos aqui referidos, cada grupo é identificado e seguido por um trilho comum a todos os elementos a ele associados. Porém, é necessário criar um mecanismo que possibilite detectar a separação de elementos de um objecto, seguindo-os separadamente ao longo do seu percurso pela área sob monitorização. Esta avaliação da necessidade de geração de novos trilhos, a partir de um trilho existente, é efectuada através da análise da *Máscara de Probabilidade* do *Modelo de Aparência* de cada trilho.

A *Máscara de Probabilidade*, por encerrar informação sobre a forma mais provável que um objecto pode assumir, possui características que se mostram apropriadas na detecção de trajectórias divergentes protagonizadas por elementos de um grupo. A separação de elementos associados a um único trilho, e por conseguinte caracterizados por um mesmo *Modelo de Aparência*, reflecte-se na actualização do *Modelo de Aparência* do grupo.

Durante a separação dos elementos, a operação de alinhamento (Secção 5.1.2) associa o trilho ao elemento que garante um valor óptimo para a função de ajuste. Este ajustamento do trilho a um dos elementos provoca na *Máscara de Probabilidade* o desvanecimento das formas dos restantes elementos do grupo.

Figura 5.4. Sequência de *Máscaras de Probabilidade* de um grupo em desagregação, acompanhadas (à direita de cada *MP*) pela identificação das áreas de maior probabilidade.

Como se pode verificar na Figura 5.4, a desagregação de um grupo tem como efeito na *Máscara de Probabilidade* a divisão da forma do objecto em múltiplas regiões. Através da análise destas regiões é possível identificar o surgimento de novos trilhos. Por conseguinte, decomposição de um trilho é executada sempre que na *Máscara de Probabilidade* os pontos de maior probabilidade originarem mais que uma região, e cujas áreas sejam superiores à área mínima permitida para um objecto.

Quando verificadas as condições de desagregação o *Modelo de Aparência* do trilho em análise é actualizado, sendo-lhe atribuída a região da *Máscara de Probabilidade* que maximizou a função de ajuste. A *Imagem de Aparência* é também adaptada de modo a reflectir esta transformação. Para cada uma das restantes regiões divergentes é gerado um novo trilho com identificador distinto.

5.2. Classificação de Objectos

Num sistema de vídeo-vigilância é indispensável distinguir os diversos tipos de objectos que habitualmente povoam a área monitorizada. Tal imposição surge pela existência de características que são intrínsecas a determinadas classes de objectos. Como exemplo, as pessoas tendem a deslocar-se a uma velocidade distinta da usualmente observada nos veículos. Adicionalmente, observam-se também divergências entre os percursos próprios a pessoas e a veículos.

Tendo como finalidade a detecção e previsão de eventos de quebra de segurança em vídeo-vigilância, de entre a multiplicidade de objectos tipicamente observados em espaços públicos, constituem alvos de interesse as pessoas, os grupos de pessoas e os veículos. Com este propósito, torna-se necessário desenvolver uma técnica que permita realizar a categorização dos objectos em três classes: "pessoa", "grupo" e "veículo".

Os *Modelos de Aparência*, gerados e mantidos pelo processo de seguimento de objectos, reúnem um conjunto de dados que definem a sua forma e cor. Um factor valorizado nos *Modelos de Aparência* é a sua estabilidade. Ao contrário do que acontece com as variações bruscas que ocorrem numa sequência de regiões segmentadas de um objecto, a informação contida nos *Modelos de Aparência* tendem a suavizar tais variações, proporcionando um modelo estabilizado da sua forma. Pela análise destes dados deve ser possível à técnica de classificação, identificar o tipo de objecto observado.

Existem no entanto algumas condicionantes que se devem considerar. Em aplicações de vídeo-vigilância, o equipamento de aquisição de imagem encontra-se geralmente posicionado de modo a permitir um campo de observação suficientemente amplo. A distância entre a câmara de vídeo e a área a monitorizar deve ser seleccionada de forma a maximizar a área observável, garantindo ao mesmo tempo a aquisição de características diferenciadoras dos objectos.

Com um campo de visão alargado, informações próprias a determinados tipos de objectos deixam de ser adquiridas. Por exemplo, a identificação de pessoas através do reconhecimento de regiões com tons de pele [Moreno et al., 2001; Ghidary et al., 2000] torna-se inexequível em objectos representados por uma imagem de, por exemplo, 10 por 22 pontos. As características associadas à cor perdem assim relevância no contexto da classificação de objectos, quando estes se encontram afastados do ponto de observação.

A forma dos objectos, representada na *Máscara de Probabilidade* de cada trilho, constitui uma fonte de informação com realce para a sua classificação. Outras características discriminativas podem no entanto ser utilizadas. A análise da dispersão da região segmentada de um objecto [Lipton et al., 1998], a razão entre a sua altura e largura [Senior et al., 2001], e a razão entre a área da região de interesse pela área do objecto [Wang & Lin, 2003] devem igualmente ser apreciadas.

Pela avaliação efectuada em [Wang & Lin, 2003] sobre a eficácia destas características na discriminação de pessoas e veículos, verifica-se que a dispersão permite uma robusta detecção de veículos, embora a classificação de pessoas seja praticamente aleatória (58.7% de taxa de acerto). Por outro lado, quer a razão entre a área de interesse pela área do objecto, quer a razão entre a sua altura e largura, possibilitam taxas globais de acerto muito superiores (91.7% e 99.4% respectivamente). Tendo em consideração estes resultados, a detecção de pessoas deve basear-se nos valores obtidos através do cálculo da razão entre a altura e a largura da região definida pela *Máscara de Probabilidade* do objecto.

O primeiro passo para a classificação de objectos consiste na selecção de uma região pertencente à *Máscara de Probabilidade* que defina a forma mais provável do objecto. Para tal, seleccionam-se todos os pontos da máscara, cuja probabilidade seja superior a 50%. O conjunto desses pontos forma a *Máscara de Classificação* (*MC*):

$$MC(x,y) = \begin{cases} 1 & , \text{ se } \quad MP(x,y) \geq 0.5 \\ 0 & , \text{ caso contrário} \end{cases} \tag{5.21}$$

Utilizando a *Máscara de Classificação*, prossegue-se com o cálculo da área, altura e largura da nova região. Com estes dados, é já possível efectuar uma primeira triagem dos objectos. Assim, recorrendo a uma característica da forma dos seres humanos (i.e. a relação entre a largura e a altura de um indivíduo), é possível detectar pessoas que se deslocam isoladamente. Para o efeito, define-se que um determinado objecto pertence à classe "pessoa" se a razão entre a largura e a altura desse objecto for igual ou inferior a 0.8, ou seja, se a largura não for superior a 80% da altura do objecto:

$$tipo_k = "pessoa" \quad , \; se \quad \frac{largura}{altura} \leq 0.8 \tag{5.22}$$

Contudo, a discriminação entre veículos e grupos de pessoas não é viável segundo esta abordagem. Não se verifica uma diferenciação aceitável entre aqueles rácios nas classes de

veículos e de grupos. Torna-se assim necessário identificar uma nova característica discriminativa entre estas classes. Em ambos os casos a largura ultrapassa 80% da altura do objecto. Assim, o passo seguinte consiste na concepção dos histogramas horizontais da *Máscara de Classificação*. Estes histogramas fornecem um meio de diferenciação entre as duas classes remanescentes.

Observe-se os seguintes histogramas horizontais apresentados na Figura 5.5. Através da análise dos histogramas, verifica-se que existe uma diferença significativa na forma destes, quando se comparam objectos rígidos, como os veículos, com grupos de pessoas. Isto é, os grupos geram frequentemente histogramas com oscilações bastante acentuadas que indicam as zonas de separação dos indivíduos pertencentes ao grupo.

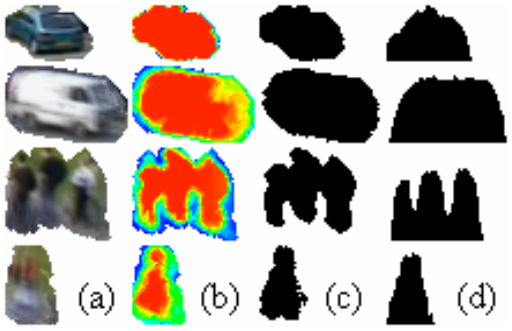


Figura 5.5. (a) *Imagens de Aparência*; (b) *Máscaras de Probabilidade*; (c) *Máscaras de Classificação*; (d) Histogramas horizontais.

Explorando este facto, é possível definir um método que permita distinguir veículos de grupos de pessoas. A técnica adoptada consiste, em encontrar duas regiões cujas áreas sejam superiores a 1/8 da área total do histograma. Essas regiões são definidas através da intersecção das zonas resultantes da parte superior do histograma, dissecado por um eixo horizontal, que se inicia na base e atinge metade da altura máxima do objecto. Sempre que se detectem duas regiões com área superior a 1/8 da área total do histograma, classifica-se o objecto como pertencente à classe de "grupo", caso contrário trata-se de um veículo.

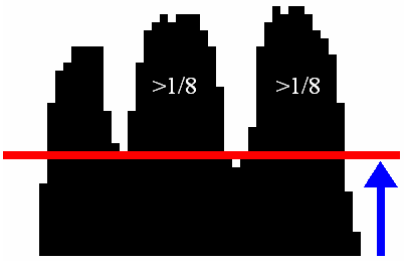


Figura 5.6. Exemplo do processo de classificação de um grupo de pessoas.

5.3. Avaliação da Técnica de Seguimento de Objectos

De modo a avaliar o sistema de segmentação e seguimento de objectos aqui proposto, recorreu-se ao conjunto de dados de referência disponibilizados pela [PETS, 2001b]. Tirando partido da natureza pública e gratuita deste conjunto de dados, a sua utilização na avaliação do sistema permite não só efectuar uma análise comparativa com os sistemas previamente testados sobre a mesma sequência de vídeo, mas possibilitará também uma comparação com sistemas de seguimento de objectos que venham a ser desenvolvidos.

Com efeito, para além dos dados de referência, é ainda necessário identificar e definir um conjunto de métricas que permitam avaliar quantitativamente o desempenho do sistema. Todavia, a escolha das métricas é dependente do tipo de dados de referência disponibilizados, tornando-se necessário realizar uma apreciação prévia da sua natureza.

5.3.1. Esquema *XML* dos Dados de Referência

No âmbito das séries de *workshops* da *PETS – Performance Evaluation of Tracking and Surveillance* foi proposta uma sequência de imagens de teste, com o objectivo de possibilitar a avaliação do desempenho dos sistemas de seguimento de objectos sobre uma plataforma comum. Deste esforço de padronização resultou também uma especificação de esquema *XML* para representação dos resultados do seguimento de objectos, bem como um conjunto de dados de referência nesse formato.

No esquema *XML* proposto, uma sequência é apresentada como um conjunto de imagens de vídeo, onde para cada imagem são descritos os objectos detectados. Estes são então representados pela informação bidimensional e tridimensional que os caracteriza. Na Figura 5.7 é possível visualizar o esquema *XML* utilizado para a representação da informação resultante da segmentação e seguimento de objectos.

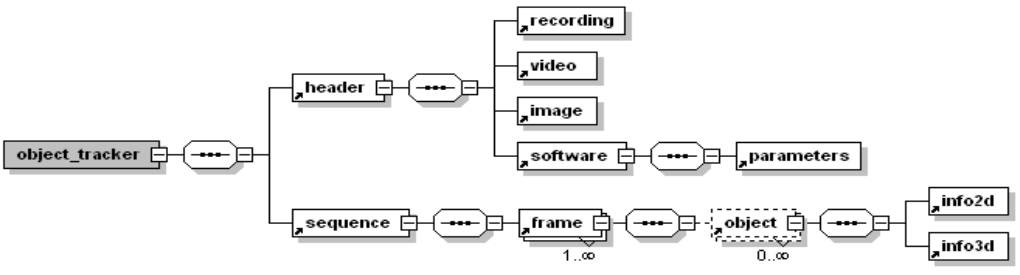


Figura 5.7. Esquema *XML* para os dados de referência adoptado de [Young & Ferryman, 2005].

A cada imagem da sequência descrita (*frame* no esquema *XML*) é-lhe atribuído um identificador, acompanhado pela indicação do número de objectos segmentados. Estes objectos são caracterizados por um identificador exclusivo, que é mantido para todas as imagens da sequência em que o mesmo seja detectado. Embora a especificação proposta pela *PETS* possibilite a descrição tridimensional de cada objecto, apenas foi disponibilizada a descrição bidimensional. Deste modo, um objecto é representado pelo centro-de-massa e pela caixa delimitadora que o encerra.

De forma a permitir a comparação entre os dados de referência e os dados gerados pelo sistema de seguimento proposto, é indispensável que o sistema adopte o referido esquema *XML* para a representação das trajectórias dos objectos numa sequência de vídeo. Conjuntamente, torna-se ainda necessário efectuar uma correcta associação entre os objectos de referência e os identificados pelo teste do sistema.

5.3.2. Métricas de Desempenho

A avaliação do desempenho de um sistema de seguimento de objectos é executada através de métricas baseadas na imagem e nos objectos. As **métricas baseadas na imagem** analisam o desempenho do processo de segmentação para cada imagem da sequência, descurando a identidade dos objectos. Nesta abordagem, cada imagem é sujeita à verificação do número de objectos segmentados, analisando a sua área e localização, relativamente aos dados de referência. Os resultados são apresentados como uma média das medições efectuadas para a totalidade das imagens que compõem a sequência. Por outro lado, na avaliação de desempenho por **métricas baseadas nos objectos**, considera-se a totalidade da trajectória de cada objecto na sequência em análise. O recurso a este tipo de métricas, onde cada trilho é encarado como uma entidade independente, permite analisar o desempenho do processo de seguimento de objectos.

No âmbito das **métricas baseadas na imagem**, para cada imagem da sequência calcula-se um conjunto de quatro variáveis de suporte:

– **Verdadeiro Positivo** (*VP*): Número de imagens onde o resultado da segmentação está em concordância com os dados de referência na existência de um ou mais objectos, e onde pelo menos uma região segmentada coincide com os objectos de referência.

- **Verdadeiro Negativo** (*VN*): Número de imagens onde o resultado da segmentação está em concordância com os dados de referência na ausência de qualquer objecto.

- **Falso Positivo** (*FP*): Número de imagens onde o resultado da segmentação indica a presença de pelo menos um objecto, enquanto os dados de referência não apresentam qualquer objecto, ou apresentando, não coincidem com as caixas delimitadoras das regiões segmentadas.

- **Falso Negativo** (*FN*): Número de imagens onde os dados de referência apresentam pelo menos um objecto, enquanto o resultado da segmentação não apresenta qualquer região, ou existindo, não coincidem com as caixas delimitadoras dos objectos identificados pelos dados de referência.

Uma região segmentada diz-se coincidente com um objecto de referência se: o centro-de-massa da região segmentada se situar no interior da caixa delimitadora do objecto de referência; ou se o centro-de-massa do objecto de referência residir no interior da caixa delimitadora da região segmentada. Define-se ainda *TO* como o número total de imagens para as quais os dados de referência identificam a presença de objectos, e *TI* como o número total de imagens da sequência de vídeo.

A avaliação do desempenho por **métricas baseadas nos objectos** atribui um novo significado às variáveis de suporte (*VP*, *VN*, *FP*, *FN*). A análise de uma técnica de seguimento de objectos implica a identificação de cada objecto em movimento ao longo de toda a sequência. Assim, o foco de interesse passa para os trilhos gerados pelos objectos. Deste modo, as variáveis de suporte passam a ser calculadas para cada trilho de referência, sendo-lhes atribuídos os seguintes significados:

- **Verdadeiro Positivo** (*VP*): Número de imagens onde o resultado da técnica de seguimento exibe um trilho coincidente com os dados de referência.

- **Verdadeiro Negativo** (*VN*): Número de imagens onde o resultado da técnica de seguimento está em concordância com os dados de referência na ausência de qualquer trilho detectado.

- **Falso Positivo** (*FP*): Número de imagens onde o resultado da técnica de seguimento indica a presença de um trilho não coincidente com os dados de referência.

– **Falso Negativo** (*FN*): Número de imagens onde os dados de referência apresentam um trilho não coincidente com o resultado da técnica de seguimento.

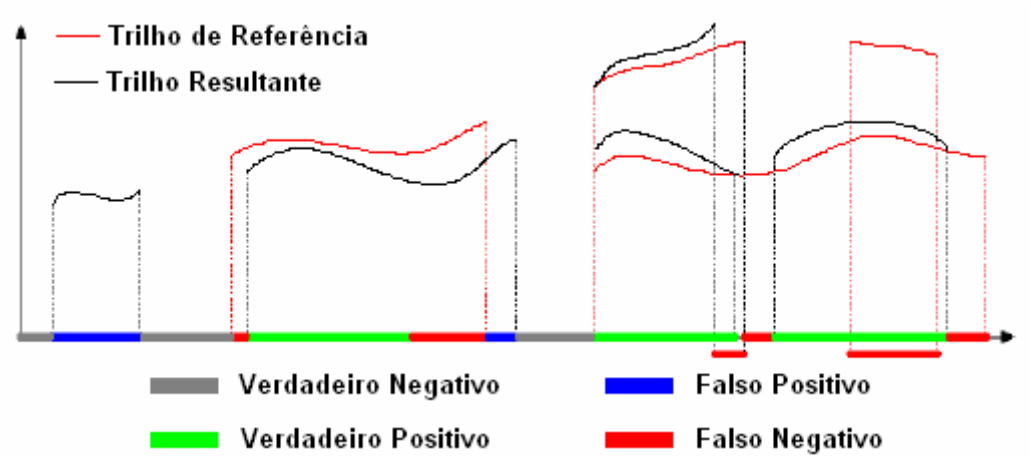


Figura 5.8. Exemplo de diferentes possibilidades da associação de trilhos obtidos pela técnica de segmentação, com os trilhos de referência.

Após o cálculo das variáveis de suporte, no contexto das métricas baseadas nas imagens e nos objectos, a avaliação da técnica de segmentação e seguimento de objectos prossegue com a computação das seguintes métricas de desempenho [Bashir & Porikli, 2006]:

$$Taxa\ de\ Detecção\ do\ Sistema\ de\ Seguimento\ (TDSS) = \frac{VP}{TO} \tag{5.23}$$

$$Taxa\ de\ Falso\ Alarme\ (TFA) = \frac{VP}{TP + VP} \tag{5.24}$$

$$Taxa\ de\ Detecção\ (TD) = \frac{VP}{VP + FN} \tag{5.25}$$

$$Especificidade = \frac{VN}{FP + VN} \tag{5.26}$$

$$Exactidão = \frac{VP + VN}{TI} \tag{5.27}$$

$$Taxa\ de\ Falsos\ Negativos\ (TFN) = \frac{FN}{FN + VP} \tag{5.28}$$

$$Taxa\ de\ Falsos\ Positivos\ (TFP) = \frac{FP}{FP + VN} \tag{5.29}$$

5.3.3. Características do Sistema de Teste

O sistema de detecção e seguimento de objectos foi implementado em linguagem de programação C, sobre um computador pessoal padrão, baseado num processador de 32 *bits* a 3.0GHz, com 512MB de memória RAM, e correndo um sistema operativo *GNU/Linux*.

A escolha da linguagem C como forma de implementar o conjunto de técnicas desenvolvidas foi baseada na rapidez de execução quando comparada com outras linguagens de programação (e.g. Java). A aptidão para a optimização do código bem como o fácil manuseamento de memória foram igualmente factores que pesaram nesta decisão.

A utilização de um sistema operativo aberto, em detrimento de uma solução proprietária, tem por base a intenção de construir um protótipo que possa vir a ser facilmente incorporado nas mais recentes soluções de *CCTV*. Actualmente, verifica-se que o mercado da vídeo-vigilância caminha para a descentralização do processamento. Na nova filosofia, as câmaras de vídeo encontram-se dotadas de capacidade de processamento, e utilizam sistemas operativos baseados em *Linux* [Regazzoni et al., 2001; Bramberger et al., 2006].

5.3.4. Resultados Experimentais

A avaliação do tempo de execução da técnica de seguimento de objectos foi realizada através de seis testes sobre a sequência de vídeo. Em cada teste, procedeu-se às medições do tempo de processamento requerido para cada imagem da sequência. Deste modo, o tempo médio de execução é obtido pela média aritmética da totalidade dos testes efectuados para cada imagem.

Realizando várias medições do tempo de execução, é possível obter um valor de maior confiança, reduzindo o efeito de perturbações com origem em rotinas e aplicações próprias ao sistema operativo. Com efeito, não se consideram para análise os resultados que apresentem um forte desvio em relação às restantes medições.

Na Figura 5.9 e Figura 5.10 apresentam-se os resultados obtidos na avaliação do tempo de execução da técnica de seguimento de objectos. Estes resultados são acompanhados pelos os valores auferidos sobre a área segmentada para cada imagem (Figura 5.9), bem como a área definida pela totalidade dos *Modelos de Aparência* (Figura 5.10). Com este conjunto de medições é então exequível a realização de uma avaliação de correlação entre o tempo necessário para o processo de seguimento com estes dois factores.

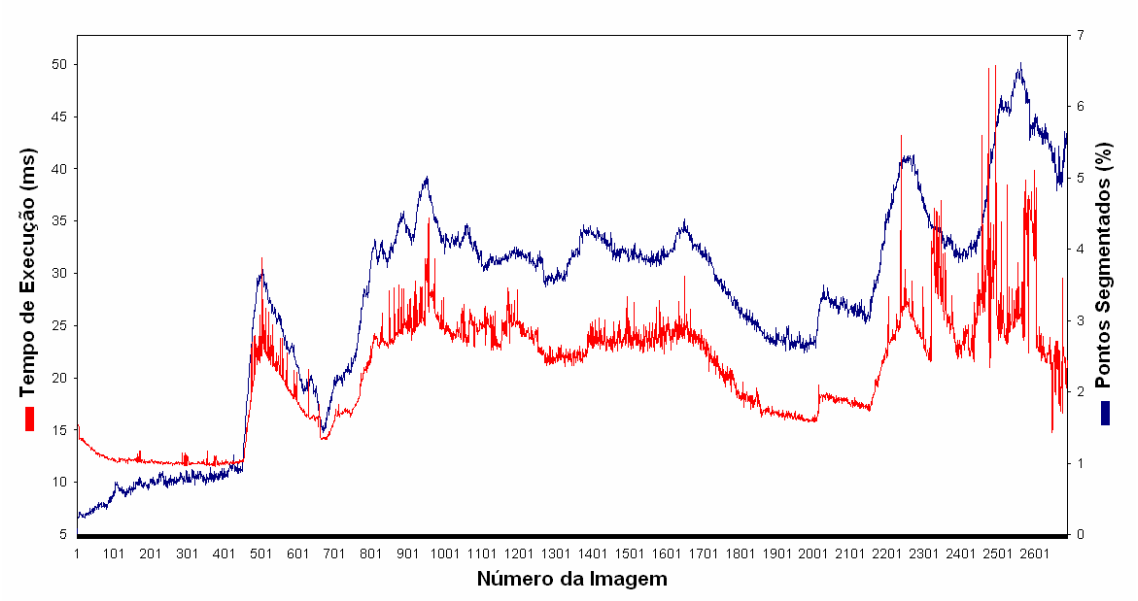


Figura 5.9. Tempo de execução da técnica de seguimento de objectos, acompanhado pela percentagem de pontos segmentados por imagem da sequência.

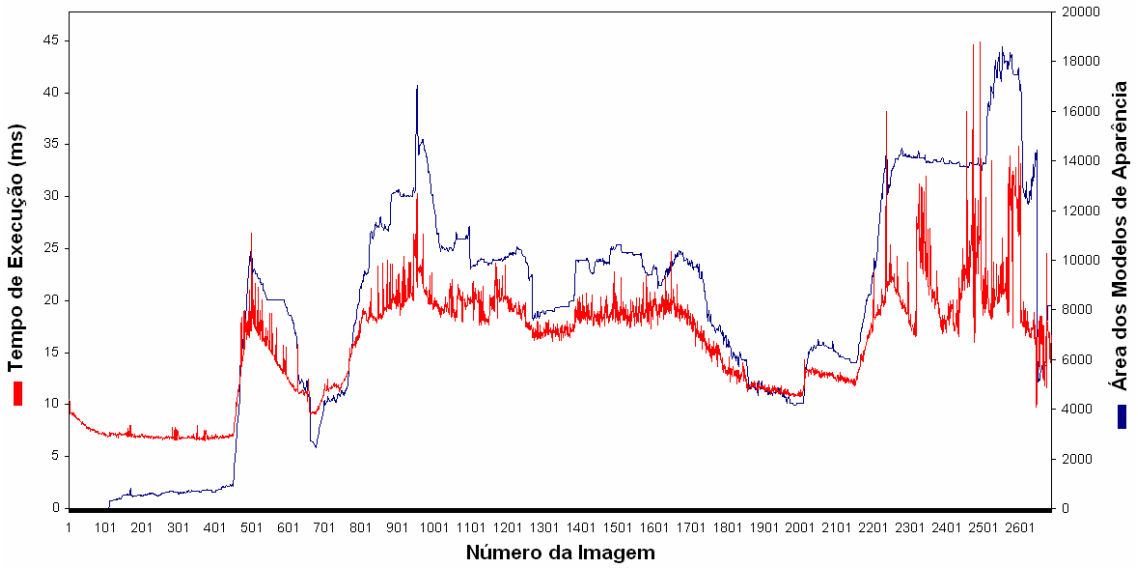


Figura 5.10. Tempo de execução da técnica de seguimento de objectos, acompanhado pela área dos *Modelos de Aparência*.

Com base nestes resultados, obteve-se um valor de tempo médio de execução da técnica de seguimento de objectos de 15619µs por imagem, com um desvio padrão de 5442µs. Como se pode verificar, existe uma forte variação do tempo de execução. Os testes efectuados demonstraram ainda um elevado coeficiente de correlação (0.88) entre o tempo de execução da técnica de seguimento de objectos, com o número de pontos segmentados em cada imagem. Contudo, verificou-se que existe uma correlação superior entre o tempo de execução e a área definida pelos *Modelos de Aparência* (0.92).

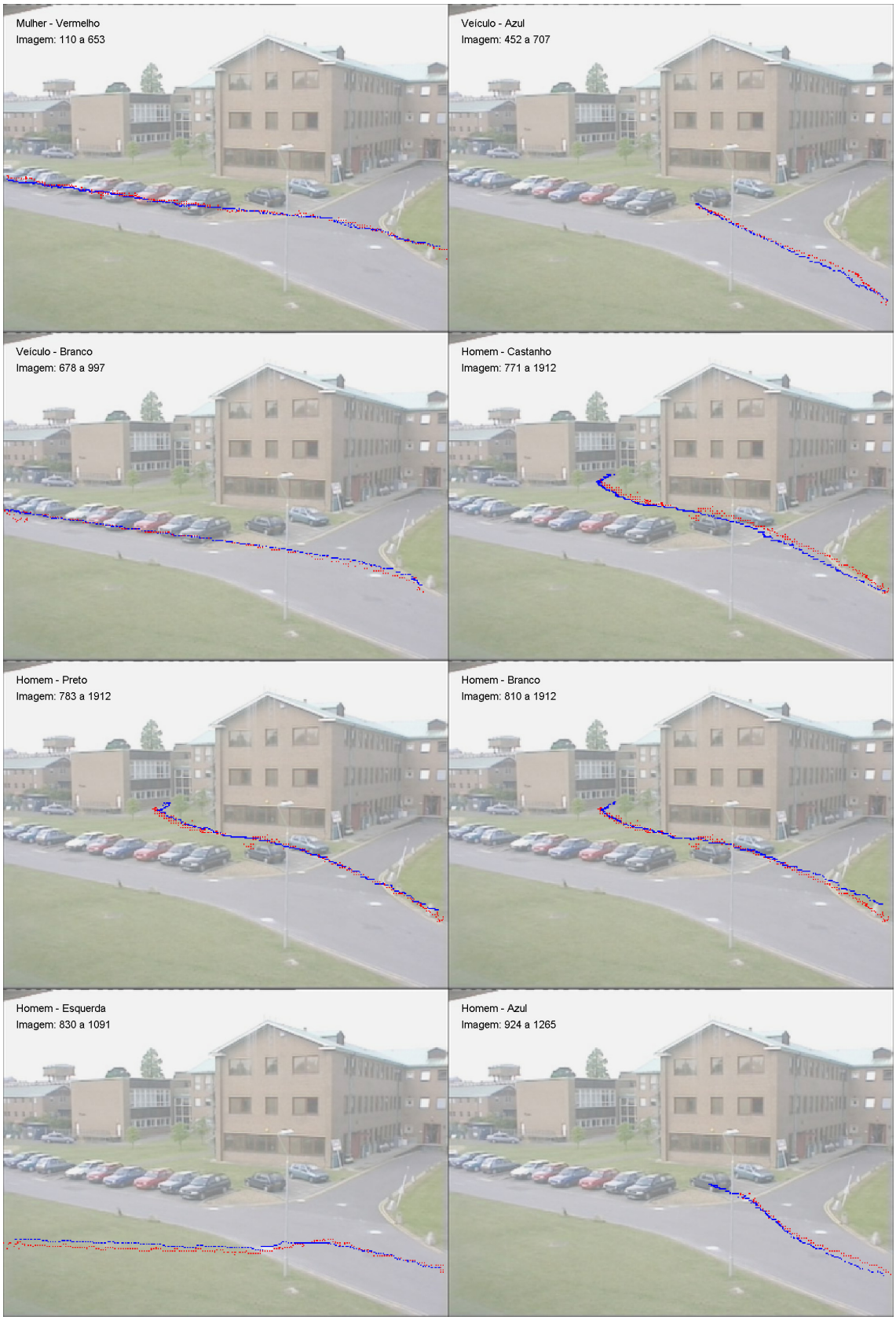


Figura 5.11. Alguns trilhos gerados pelo sistema de segmentação (a vermelho) e respectivos trilhos de referência (a azul), sobrepostos à imagem de plano de fundo do conjunto de imagens de teste.

O resultado dos processos de segmentação e seguimento de objectos em movimento pode ser apreciado na Figura 5.11. Como se pode verificar, os trilhos gerados pelo sistema proposto apresentam uma elevada similaridade com os trilhos de referência. Destaca-se sobretudo a integridade do sistema de seguimento de objectos aquando da oclusão parcial, pelo cruzamento de diversos objectos em movimento.

Após a medição das variáveis de suporte, quer para as métricas baseadas na imagem, quer para as métricas baseadas nos objectos, obtiveram-se os resultados apresentados na Tabela 5.3. O sistema proposto exibe, para as métricas baseadas na imagem, valores muito próximos dos ideais. Como exemplo, destaca-se o desempenho auferido pela *Taxa de Detecção do Sistema de Seguimento* (*TDSS*), pela *Taxa de Falsos Alarmes* (*TFA*), pela *Taxa de Detecção* (*TD*), *Exactidão*, e *Taxa de Falsos Negativos* (*TFN*).

As métricas baseadas nos objectos apresentam resultados com um maior afastamento dos valores ideais, sendo que a *Taxa de Falsos Positivos* (*TFP*) evidencia o mais relevante decréscimo de desempenho. A explicação para esta ocorrência deve-se sobretudo ao facto do sistema proposto manter o trilho dos objectos por um período de tempo superior ao apresentado nos dados de referência. Tal facto manifesta-se principalmente nos trilhos de objectos que interrompem o seu movimento de deslocação. Apesar da influência que os *FP* mantêm sobre algumas das métricas, a *Taxa de Detecção do Sistema de Seguimento* (*TDSS*), a *Taxa de Detecção* (*TD*), e a *Exactidão*, manifestam no entanto valores bastante satisfatórios.

Tabela 5.3. Resultados da avaliação de desempenho do sistema proposto, através de métricas baseadas na imagem e nos objectos, sobre o conjunto de imagens de teste da *PETS* 2001.

Métrica	Baseada na Imagem	Baseada nos Objectos
TDSS	0.98	0.75
TFA	0.01	0.02
TD	0.98	0.80
Especificidade	0.82	0.49
Exactidão	0.97	0.80
TFN	0.02	0.20
TFP	0.18	0.51

5.4. Discussão

O seguimento de objectos em movimento, observados a partir de imagens coloridas e digitalizadas, mostrou ser viável através do recurso a *Modelos de Aparência*. Com a abordagem proposta, cada objecto é modelado pela sua forma e cor. A informação discriminativa de um objecto (*Imagem de Aparência* e *Máscara de Probabilidade*) é mantida em memória pelo período que este permaneça em cena.

Apesar da elevada complexidade computacional do método proposto, a sua implementação demonstrou conformidade com os requisitos de tempo-real idealizados para o sistema. A soma do tempo médio dispendido pela técnica de segmentação, com o tempo necessário à execução da técnica de seguimento perfaz um total de aproximadamente 50ms por imagem. Em tais condições, o sistema é capaz de processar cerca de vinte imagens por segundo.

A técnica proposta apresenta contudo variações relevantes do tempo de processamento, mostrando-se dependente das dimensões dos *Modelos de Aparência*. Por conseguinte, é de esperar uma variação da cadência de processamento das imagens, de acordo com o número e dimensões dos objectos em cena.

Se se comparar o resultado das métricas de desempenho do sistema proposto, com os valores obtidos para o "*Multi-kernel Meanshift Tracking System*" e para o "*Ensemble Tracking System*" [Bashir & Porikli, 2006], verifica-se que, no que respeita às métricas baseadas na imagem, o sistema aqui proposto apresenta um desempenho consideravelmente superior em todas as métricas analisadas. Esta supremacia é também evidenciada nas métricas baseadas nos objectos, à excepção da *Especificidade* e da *Taxa de Falsos Positivos* (*TFP*). Estas métricas apresentam um desempenho inferior ao exibido pelos sistemas analisados, devido à influência exercida pelos *Falsos Positivos* (*FP*).

De acordo com o referido na secção anterior, a presença de um número significativo de *FP* deve-se sobretudo à manutenção dos trilhos dos objectos por um período de tempo superior ao apresentado nos dados de referência. O que de facto se verifica é a manutenção do seguimento de objectos que, após um período de movimento, se encontram imóveis na cena monitorizada. Como tal, este valor de *FP* não deve ser encarado como uma influência negativa no sistema proposto, uma vez que não constitui uma fonte de erro no seguimento de objectos.

Tabela 5.4. Comparação dos resultados obtidos com duas técnicas (*Multi-kernel Meanshift Tracking System* e *Ensemble Tracking System*) avaliadas em [Bashir & Porikli, 2006] (melhor resultado a negrito).

Métrica	Baseada na Imagem			Baseada nos Objectos		
	OBSERVER[17]	MMTS[18]	ETS[19]	OBSERVER	MMTS	ETS
TDSS	**0.98**	0.54	0.61	**0.75**	0.68	0.69
TFA	**0.01**	0.21	0.21	**0.02**	0.21	0.22
TD	**0.98**	0.54	0.61	**0.80**	0.65	**0.80**
Especificidade	**0.82**	0.58	0.60	0.49	0.64	**0.68**
Exactidão	**0.97**	0.59	0.64	0.80	0.80	**0.83**
TFN	**0.02**	0.45	0.39	0.20	0.34	**0.19**
TFP	**0.18**	0.41	0.40	0.51	0.36	**0.32**

[17] OBSERVER foi a denominação atribuída ao sistema desenvolvido no âmbito deste trabalho de doutoramento.

[18] Multi-kernel Meanshift Tracking System [Bashir & Porikli, 2006].

[19] Ensemble Tracking System [Bashir & Porikli, 2006].

Capítulo 6

6. Detecção e Previsão de Comportamentos

Este capítulo é dedicado à detecção e previsão automática de comportamentos, considerados não usuais e anormais, recorrendo a técnicas de inteligência artificial. Como tal, apresentam-se dois classificadores (*N-ary Trees* e *Dynamic Oriented Graph*) propostos no âmbito deste trabalho e que, utilizando os dados provenientes das funções de processamento e análise de imagem, permitem modelar sequências temporais.

Inicia-se o presente capítulo com a apresentação da abordagem adoptada para a classificação de comportamentos, onde se pretende limitar ao mínimo indispensável a utilização de informação de contexto. Prossegue-se com a descrição do *N-ary Trees*, um classificador não supervisionado, que através de uma fase de treino adquire um modelo dos padrões observados. Seguidamente é apresentado o *Dynamic Oriented Graph (DOG)*, um classificador que implementa o mesmo conceito que o anterior, mas cuja aprendizagem, igualmente não supervisionada, é realizada de modo a permitir o ajuste dinâmico do modelo, de acordo com as alterações de padrões observadas ao longo do tempo.

Ainda neste capítulo, os classificadores propostos são avaliados recorrendo a um conjunto de sequências temporais geradas de forma artificial. O classificador com melhor desempenho (*DOG*), foi implementado no protótipo do sistema e testado em ambiente real.

6.1. Abordagem para o Classificador de Comportamentos

Uma corrente usual na análise e reconhecimento de comportamentos em vídeo-vigilância advoga que o comportamento de um objecto pode ser descrito por uma sequência de acções atómicas (e.g. andar, correr, parar) que este executa num determinado ambiente [Dempster et al., 1977; Naylor, 2006]. Tal abordagem requer geralmente o conhecimento do histórico de estados dos atributos de maior relevância e representatividade do objecto alvo (e.g. mãos, pés e cabeça). A aquisição de medições destes atributos pode ser difícil e em muitos casos impraticável, quando se pretendem monitorizar situações reais em ambientes não-laboratoriais.

Em instalações de vídeo-vigilância, os objectos encontram-se frequentemente afastados da câmara de vídeo e a existência de vários objectos em simultâneo, movimentando-se na cena, originam fenómenos problemáticos (e.g. oclusão) que podem induzir erros ou falhas na detecção. Uma outra adversidade advém do facto de, apesar da existência de métodos bem conhecidos para a identificação de acções atómicas, o reconhecimento do início e fim de tais eventos é de extrema dificuldade. Um obstáculo suplementar consiste na procura pela descrição da correcta sequência de acções, a qual é principalmente baseada no conhecimento que o utilizador detém sobre os comportamentos de interesse.

Outra proposta para a análise de comportamento, em vez de se fixar sobre uma cadeia de acções atómicas, analisa a totalidade da sequência temporal com o intuito de assimilar e reunir padrões de actividades distintas [Mecocci et al., 2003]. Este género de abordagem pode ser considerada mais robusta, dado que os comportamentos são extrapolados numa perspectiva global e não por uma restritiva sequência de acções. Tal abordagem requer, no entanto, sofisticados métodos de mineração (*data mining*) para elevados volumes de dados, com o objectivo de descobrir padrões comuns a determinados tipos de comportamentos. A informação resultante deste processo é então utilizada na construção de modelos que representam tais comportamentos.

Na classificação de comportamentos é usual o recurso a informação de contexto, quer seja sobre a actividade ou sobre o ambiente monitorizado, como forma de melhorar as capacidades da técnica de classificação. Como exemplo, no âmbito do projecto ADVISOR [Naylor, 2006], a detecção de actos de vandalismo implica a modelação tridimensional do espaço monitorizado, bem como a descrição dos eventos relevantes (efectuada por peritos

de segurança através de uma linguagem de descrição especialmente concebida para o efeito).

No âmbito deste trabalho pretende-se desenvolver um classificador que reduza ao mínimo a utilização de informação de contexto. Tal opção de concepção baseia-se numa preferência pela capacidade de generalização do classificador em detrimento da precisão e detalhe da tipologia dos eventos. Num sistema de vigilância, salvo aplicações específicas, não é viável a descrição da totalidade das actividades de interesse. Em ambientes reais, o número de actividades é elevado, as sequências de acções são complexas, e alterações no ambiente implicam uma redefinição das actividades suspeitas.

Da perspectiva da segurança os comportamentos podem ser coligidos em três tipos base: **normal**, **não usual** e **anormal**. Tipicamente, numa área sob vigilância o foco de atenção recai na detecção de **comportamentos não usuais** (e.g. uma pessoa a correr no *lobby* de um hotel) e na identificação de **comportamentos anormais** (e.g. violação de uma área restrita ou uma pessoa atravessando uma via ferroviária).

Como comportamentos **normais** entendem-se aqueles que são frequentemente observados, não originando a violação de qualquer área restrita. Eventos **não usuais** são aqueles cuja frequência de acontecimento é bastante reduzida ou que a ocorrência nunca foi verificada. Quando uma acção leva à violação de uma área restrita, deve então ser classificada como um comportamento **anormal**.

Com a adopção desta abordagem, não existe necessidade de definir contextos complexos para cada cena ou actividade. Ao utilizador apenas é requerido que defina as regiões onde se aplicam restrições a determinados tipos de objectos (i.e. pessoas, grupos de pessoas, e veículos).

Com base nesta classificação de comportamentos, a detecção de eventos não usuais e a previsão de actividades anormais pode ser definida pelas seguintes questões: A trajectória de um objecto com propriedades específicas (como cor, área, perímetro) é reconhecida? Se é reconhecida, baseando-se no histórico da trajectória e propriedades (e.g. cor do objecto), qual é a probabilidade do objecto seguir um caminho que o levará à violação de uma área restrita?

6.2. Classificador *N-ary Trees*

Uma possível técnica para modelação de comportamentos é conseguida através do recurso a uma estrutura de nós, conectados de forma unidireccional, cada um definindo uma região no hiperespaço de atributos, para um intervalo de tempo predefinido, tendo associada uma probabilidade de geração de eventos anormais. Os dados fornecidos ao classificador incluem informação temporal e espacial acerca da trajectória do objecto, embora possam também ser examinados outros atributos, considerando características tais como a cor principal, a área ou o perímetro do objecto.

Formalmente, neste trabalho propõe-se um classificador denominado de *N-ary Trees*[20] cujos nós são compostos por dois tipos de informação:

- **Distribuição Gaussiana Multivariável** – $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$;
- **Probabilidade de Evento Anormal** – P_{ea} .

A distribuição gaussiana $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é representada por um vector contendo a média dos atributos dos objectos ($\boldsymbol{\mu}$) e por uma matriz de covariância ($\boldsymbol{\Sigma}$). De modo a simplificar o processo de cálculo, assume-se que não há correlação entre os diferentes atributos. A *Probabilidade de Evento Anormal* (P_{ea}) consiste na possibilidade de um objecto vir a violar uma área restrita, sabendo que este seguiu um determinado trajecto, de acordo com as relações temporais e espaciais do modelo. Por outras palavras, se um objecto percorrer um caminho semelhante a um padrão que se verificou ter conduzido a uma violação de área restrita, então a probabilidade de este objecto gerar um evento anormal é elevada.

Existem algumas semelhanças entre o classificador proposto (*N-ary Trees*) e as *Redes Bayesianas*. Ambos são definidos por um grafo acíclico direccionado, em que os arcos representam influência directa de um nó em outro. Contudo, nas *Redes Bayesianas* as variáveis são representadas exclusivamente pelos nós do grafo, em que cada nó indica um atributo. Tal facto já não se verifica no classificador *N-ary Trees*, onde cada nó define uma função densidade de probabilidade que considera todos os atributos à excepção do tempo.

[20] Considera-se que a estrutura do classificador se assemelha a um conjunto de árvores, conectadas entre si, em que as raízes são representadas pelos nós de entrada.

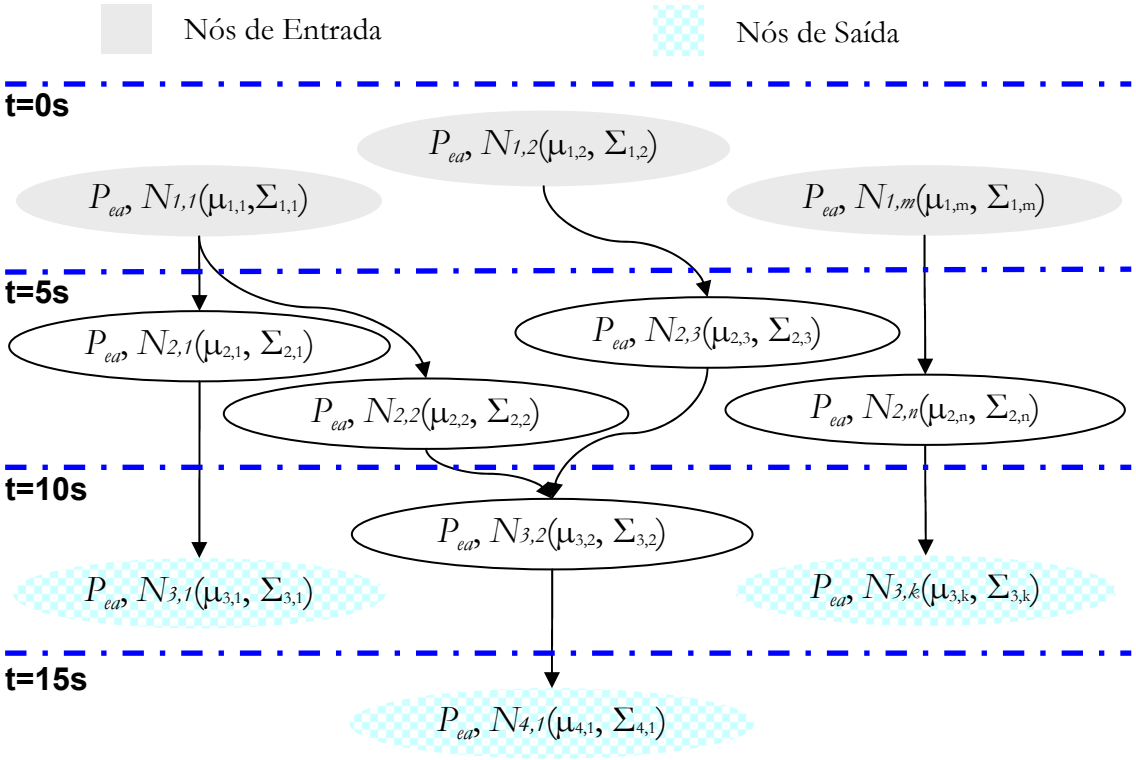


Figura 6.1. Exemplo da estrutura de um classificador *N-ary Trees*.

Como se pode verificar na Figura 6.1, o grafo de classificação encontra-se organizado por camadas temporais de tal modo que cada novo trilho deve ser descrito por uma sequência de nós (um nó por camada), principiando num nó de entrada. Em cada camada são definidas regiões no hiperespaço dos atributos, obtidas a partir de funções densidade de probabilidade do conjunto de atributos para um determinado período de tempo.

O modelo *N-ary Trees* pode então ser visto como um classificador espaciotemporal enriquecido com atributos característicos dos objectos, tais como a sua a área, perímetro, cor dominante, entre outros. Este classificador é construído a partir de trilhos previamente observados e classificados (apenas como eventos normais ou anormais), sendo definidos por sequências de vectores de atributos que descrevem as propriedades de um objecto a cada intervalo de tempo.

O primeiro passo na construção do classificador *N-ary Trees* consiste no seccionamento dos dados, i.e. trilhos, em secções de igual período temporal (e.g. secções de 250 imagens, considerando que a aquisição de imagens é efectuada em intervalos de tempos constantes). Seguidamente, é calculado o valor médio de cada atributo para cada uma das secções definidas pelo passo anterior.

A fragmentação dos trilhos em secções de intervalos temporais fixos, com a subsequente redução ao valor médio, tem como propósito a diminuição do volume de dados a tratar. Este procedimento pode ser encarado como a execução de uma amostragem da informação produzida pelo processo de análise de vídeo.

De acordo com o teorema da amostragem de Nyquist-Shannon [Shannon, 1949], o valor do período temporal deve satisfazer o critério que especifica uma frequência de amostragem superior ao dobro da mais elevada frequência observada. Transpondo este teorema para o caso em análise, pode-se afirmar que os períodos temporais das secções devem possuir um valor inferior a metade da menor flutuação dos valores dos atributos (no domínio do tempo), comum aos comportamentos de interesse.

Apesar dos fundamentos teóricos subjacentes à escolha do período temporal óptimo, no contexto deste trabalho, este é definido empiricamente. Deve ser claro que secções com períodos temporais mais curtos possibilitam uma maior precisão na construção de modelos de comportamento. Contudo, na escolha do período temporal deve-se encontrar um compromisso com a capacidade de memória disponível para o armazenamento dos modelos.

Embora se tenha optado por uma abordagem que implementa secções definidas por períodos temporais fixos, não são de excluir variantes que promovam uma utilização de secções com períodos temporais variáveis, por exemplo, em função do nível de risco de determinadas áreas sob observação.

Após a obtenção dos valores médios, prossegue-se com o agrupamento em classes destes valores para cada período de tempo, formando diferentes distribuições gaussianas. Dado não existir um conhecimento *a priori* sobre o número de classes numa camada (i.e. período temporal), é necessário recorrer a um algoritmo de *clustering* capaz de inferir sobre o número espectável de classes. Com este propósito foi utilizado um algoritmo de **Expectativa-Maximização**[21] (*EM*) [Dempster et al., 1977] com detecção automática de número de classes baseada numa técnica *k-fold Cross-validation* [Kohavi, 1995].

[21] Tradução adoptada do termo inglês: *Expectation-Maximization.*

6.2.1. Processo de Inferência das Classes

Considere-se **X** o conjunto de dados definidos por *M* vectores $\mathbf{x}_m$ de dimensão *d* (i.e. *d* atributos), onde *M* representa o número de observações que compõem o conjunto de dados para cada camada temporal.

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, ..., \mathbf{x}_M\} \text{ onde, } \quad \mathbf{x}_m = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \text{ com, } 1 \le m \le M \tag{6.1}$$

Note-se que num dado vector $\mathbf{x}_m$, cada atributo é obtido pela média dos valores observados para esse mesmo atributo no período de tempo associado à camada temporal em análise.

Assume-se que numa camada temporal os dados são gerados a partir de uma mistura de distribuições gaussianas, da qual se desconhece o número de componentes (*C*) bem como as suas propriedades, i.e. média e desvio padrão de cada distribuição. Assim, é necessário recorrer a uma técnica que permita aferir esta informação valendo-se apenas do conjunto de dados observados. Para este efeito, e como foi referido utilizou-se uma associação entre a técnica de *k-fold Cross-validation* e o algoritmo *EM*.

Inicialmente considera-se que os dados são gerados a partir de uma única distribuição gaussiana, ou seja, existe apenas uma componente (*C*=1). O número de distribuições vai sendo incrementado em uma unidade enquanto o critério de paragem não for satisfeito. Note-se que no limite o número de componentes poderá igualar o número de observações.

A técnica *k-fold Cross-validation* é aplicada então para o número de componentes definido, procedendo-se à divisão do conjunto de dados de treino em *K* subconjuntos. Para aumentar a robustez da técnica, o fraccionamento dos dados é precedido da execução de uma redistribuição aleatória dos mesmos. Após o fraccionamento, *K*-1 desses conjuntos são seleccionadas para o cálculo da *Expectativa* e posteriormente a *Maximização*, ficando o conjunto excedente reservado para a avaliação do critério de paragem. Deste modo, o teste é realizado com *M*/*K* observações, enquanto o número de observações utilizadas no treino é definido por:

$$N = \frac{M \cdot (K-1)}{K} \tag{6.2}$$

Utilizando os dados de treino, a inicialização de cada componente gaussiana é efectuada através da definição de uma probabilidade de ocorrência $P(W_c)$, sendo que à média $\boldsymbol{\mu}_c$ é-lhe atribuído um valor aleatório contido no conjunto de dados de treino. A inicialização de uma distribuição conclui com a determinação da matriz de covariância $\boldsymbol{\Sigma}_c$.

$$\begin{aligned}
&\text{Para cada} \quad c \in C \quad \text{fazer} \\
&\quad P(W_c) = \frac{1}{C} \\
&\quad \boldsymbol{\mu}_c = random(\mathbf{x}_n) \\
&\quad \boldsymbol{\Sigma}_c = \frac{\sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}}_c)^2}{N} \quad \text{, onde,} \quad \bar{\mathbf{x}}_c = \frac{\sum_{n=1}^{N} \mathbf{x}_n}{N}
\end{aligned} \tag{6.3}$$

Após a inicialização das componentes a testar, o processo segue com um ciclo de iterações de *EM*. Com o cálculo da *Expectativa* pretende-se determinar a probabilidade expectável de cada elemento do conjunto de dados observados pertencer a cada uma das distribuições:

$$\begin{aligned}
&\text{Para cada} \quad n \in N \quad \text{fazer} \\
&\quad P(\mathbf{x}_n) = \sum_{c=1}^{C} \left[P(\mathbf{x}_n \mid W_c) \cdot P(W_c) \right] \\
&\quad \text{Para cada} \quad c \in C \quad \text{fazer} \\
&\quad\quad P(\mathbf{x}_n \mid W_c) = \frac{1}{(2\pi)^{d/2} \cdot |\boldsymbol{\Sigma}_c|^{1/2}} \cdot e^{\left(-\frac{(\mathbf{x}_n - \mu_c)^T \cdot \boldsymbol{\Sigma}_c^{-1} \cdot (\mathbf{x}_n - \mu_c)}{2} \right)} \\
&\quad\quad P(W_c \mid \mathbf{x}_n) = \frac{P(\mathbf{x}_n \mid W_c) \cdot P(W_c)}{P(\mathbf{x}_n)}
\end{aligned} \tag{6.4}$$

Com a *Maximização* pretende-se aferir a máxima da verosimilhança pela maximização da probabilidade obtida na *Expectativa*. O processo de *Maximização* pode ser definido por:

$$\begin{aligned}
&\text{Para cada} \quad c \in C \quad \text{fazer} \\
&\quad \hat{P}(W_c) = \frac{\sum_{n=1}^{N} P(W_c \mid \mathbf{x}_n)}{N} \\
&\quad \hat{\boldsymbol{\mu}}_c = \frac{\sum_{n=1}^{N} \left[\mathbf{x}_n \cdot P(W_c \mid \mathbf{x}_n) \right]}{N \cdot \hat{P}(W_c)} \\
&\quad \hat{\boldsymbol{\Sigma}}_c = \frac{\sum_{n=1}^{N} \left[P(W_c \mid \mathbf{x}_n) \cdot (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_c) \cdot (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_c)^T \right]}{N \cdot \hat{P}(W_c)}
\end{aligned} \tag{6.5}$$

O ciclo *Expectativa-Maximização* termina assim que se verifique que para uma determinada iteração não ocorre uma melhoria significativa no ajustamento das componentes gaussianas ao conjunto de dados de treino. A medida de ajustamento é definida pelo *Logaritmo da Verosimilhança* sobre os dados representados pelos K-1 subconjuntos de treino.

$$\begin{aligned} l_{em} &= \prod_{n=1}^{N} P(\mathbf{x}_n) = \prod_{n=1}^{N} \left[P(\mathbf{x}_n \mid W_c) \cdot P(W_c) \right] \\ L_{em} &= \sum_{n=1}^{N} \ln\left(P(\mathbf{x}_n) \right) = \sum_{n=1}^{N} \ln\left(P(\mathbf{x}_n \mid W_c) \cdot P(W_c) \right) \end{aligned} \tag{6.6}$$

Repare-se que numa situação ideal o produto de $P(\mathbf{x}_n)$ tomaria valor unitário, i.e. a probabilidade de ocorrência do conjunto de observações seria máxima. Assim, o *Logaritmo da Verosimilhança* indica uma melhoria no ajuste do modelo aos dados observados quando o seu valor tende para zero.

Após a estabilização das componentes gaussianas, o subconjunto excluído deste processo de *Expectativa-Maximização* é utilizado para a avaliação do modelo. Esta avaliação é realizada através da estimativa do *Logaritmo da Verosimilhança* utilizando para o efeito o conjunto dos dados de teste:

$$L_k = \sum_{n=1}^{M/K} \ln\left(P(\mathbf{x}_n) \right) \tag{6.7}$$

Os processos dos cálculos da *Expectativa-Maximização* e do *Logaritmo da Verosimilhança* (L_k) são repetidos *K* vezes até que sejam obtidos os valores para cada subconjunto. O valor final do *Logaritmo da Verosimilhança* é auferido pela média dos *K* valores:

$$L = \frac{\sum_{k=1}^{K} L_k}{K} \tag{6.8}$$

Se o valor final do *Logaritmo da Verosimilhança* for inferior ao obtido para o anterior número de distribuições gaussianas, então o número óptimo de classes foi encontrado (i.e. número de classes = *C* - 1). O mesmo se verifica se com o aumento do número de classes apenas se observar um aumento residual do *Logaritmo da Verosimilhança*.

Após a determinação do número de classes numa determinada camada temporal, o cálculo das propriedades das distribuições gaussianas que modelam os dados observados é

efectuado pela aplicação do algoritmo *EM* sobre a totalidade do conjunto de dados da camada temporal em causa. O processo aqui referido é descrito pela seguinte sequência de acções:

$$\begin{aligned}&\text{fazer}\\&\qquad \text{Expectativa}\left(\mathbf{X}, P(W_c), \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)\\&\qquad \text{Maximização}\left(\mathbf{X}, P(W_c), \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)\\&\qquad L_t = \sum_{m=1}^{M} \ln\left(P(\mathbf{x}_m)\right)\\&\text{enquanto}\left(\left|L_t - L_{t-1}\right| > \varepsilon\right)\end{aligned} \tag{6.9}$$

As componentes da mistura de gaussianos, i.e. média e matriz de covariância de cada distribuição, são definidas quando a condição de paragem for satisfeita, ou seja, assim que $|L_t - L_{t-1}| < \varepsilon$.

Através da aplicação desta técnica sobre a totalidade do conjunto de dados observados, é então possível definir grupos de classes para cada camada temporal. No entanto, a simples definição das classes não possibilita *per si* o modelar das sequências temporais observadas. Assim, para gerar o classificador *N-ary Trees* é necessário estabelecer conexões entre classes de camadas adjacentes.

6.2.2. Conexão entre Classes do Classificador *N-ary Trees*

A conexão entre as diversas classes de camadas temporais contíguas, identificadas pelo processo anterior, deve reflectir os padrões de comportamento observados. A ideia consiste em estabelecer ligações entre classes (de camadas adjacentes) de modo a modelar os comportamentos, por intermédio de uma sequência de transições entre os nós do classificador.

O processo de conexão de classes é efectuado para todas as camadas temporais definidas, o que implica percorrer a totalidade das sequências observadas (trilhos), estabelecendo a cada iteração a associação entre o vector de entrada e uma classe da camada temporal correspondente. A atribuição do vector a uma das classes não é no entanto trivial, como se pode constatar na Figura 6.2. Como exemplo, veja-se a ambiguidade inerente à atribuição da amostra **x** a uma das três distribuições representadas na figura. Qual das distribuições melhor representa a amostra? A solução passa pelo estabelecimento de *funções discriminantes*.

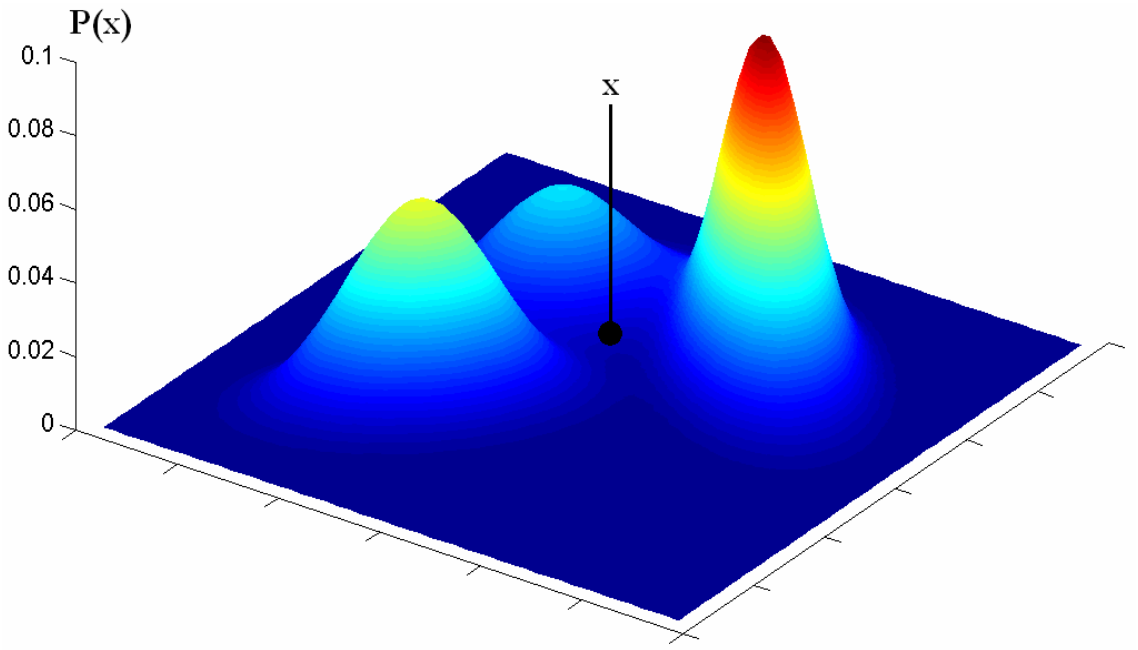


Figura 6.2. Representação gráfica de três distribuições gaussianas.

As *funções discriminantes* fornecem um meio de implementação de regras de decisão através da divisão do espaço característico em C regiões, $\mathcal{R}_1$, $\mathcal{R}_2$, $\ldots\mathcal{R}_c$. Uma forma de representação deste tipo de funções consiste em atribuir uma amostra $\mathbf{x}$ à região $\mathcal{R}_i$ que maximiza a probabilidade $P(W_i|\mathbf{x})$. Assim, define-se uma *função discriminante* $g_i(\mathbf{x})$ de acordo com a seguinte expressão [Duda et al., 2001]:

$$g_i(\mathbf{x}) = P(W_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid W_i) \cdot P(W_i)}{\sum_{c=1}^{C} P(\mathbf{x} \mid W_c) \cdot P(W_c)} \tag{6.10}$$

Como tal, se $g_i(\mathbf{x}) > g_j(\mathbf{x})$ para todo $j \neq i$, então $\mathbf{x}$ pertence a $\mathcal{R}_i$, e como resultado a regra de decisão atribui $\mathbf{x}$ à classe W_i. É no entanto possível simplificar significativamente a função discriminante devido ao facto de existir um denominador comum a todas as funções. Deste modo, a equação (6.10) pode ser reescrita como:

$$g_i(\mathbf{x}) = P(\mathbf{x} \mid W_i) \cdot P(W_i) \tag{6.11}$$

ou na forma de logaritmo natural,

$$g_i(\mathbf{x}) = \ln(P(\mathbf{x} \mid W_i)) + \ln(P(W_i)) \tag{6.12}$$

Considerando que $P(\mathbf{x} \mid W_i) \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, então a equação (6.12) é expandida para:

$$g_i(\mathbf{x}) = -\frac{1}{2} \cdot (\mathbf{x} - \boldsymbol{\mu}_i)^t \cdot \boldsymbol{\Sigma}_i^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \cdot \ln(2\pi) - \frac{1}{2} \cdot \ln(\boldsymbol{\Sigma}_i) + \ln(P(W_i)) \quad (6.13)$$

As diferentes regiões encontram-se separadas por *limites de decisão*, de tal modo que se $\mathcal{R}_i$ e $\mathcal{R}_j$ forem regiões contíguas, o limite entre elas é definido por:

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \quad (6.14)$$

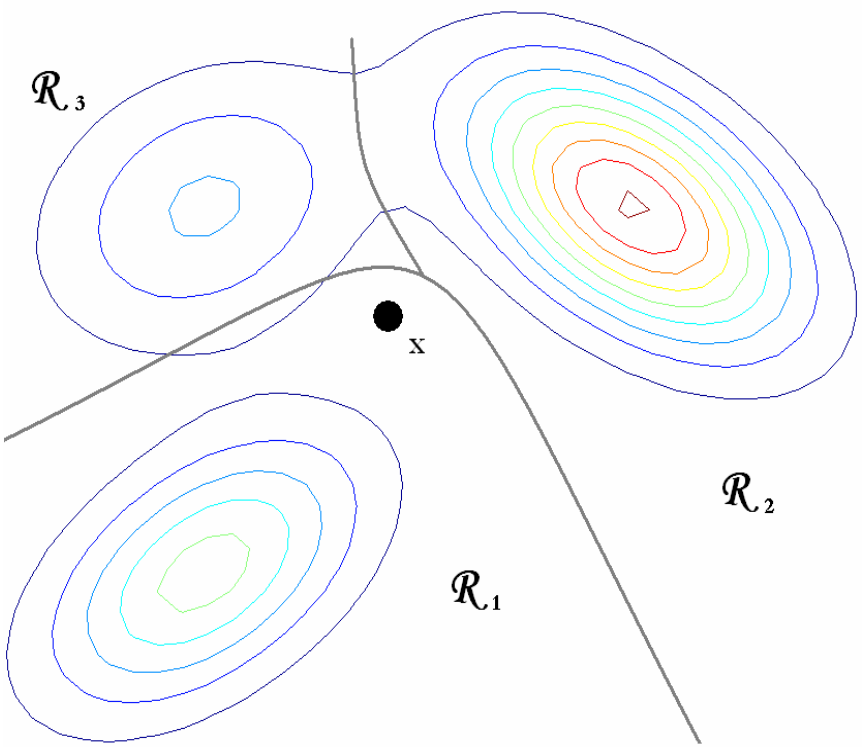


Figura 6.3. Limites de decisão entre as regiões que definem três classes.

Recorrendo a este método, como forma de realizar a atribuição das amostras apresentadas pelos trilhos de treino às classes definidas para cada camada temporal, é então possível traçar ligações entre classes de modo a definir o comportamento do objecto observado ao longo do tempo. Como resultado é obtida uma estrutura de grafo direccionado acíclico semelhante ao apresentado Figura 6.4.

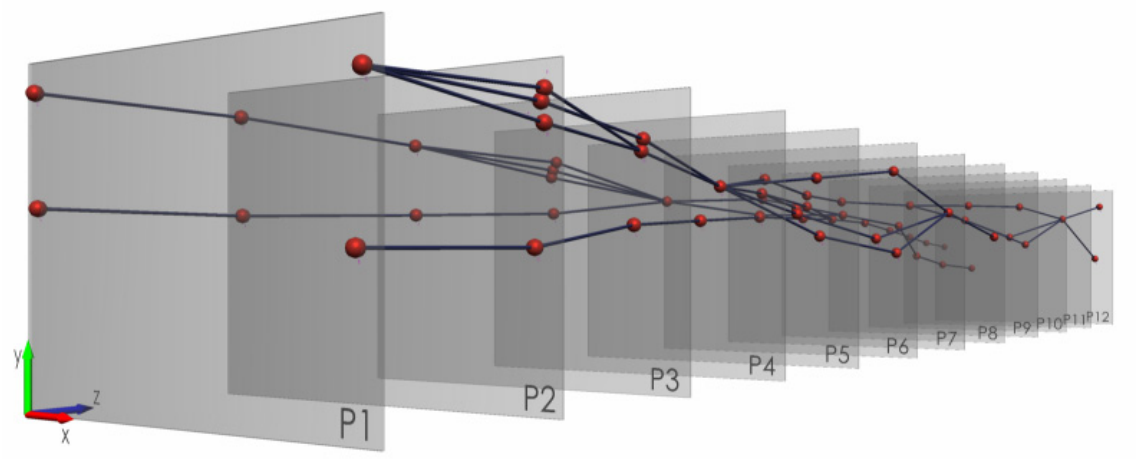


Figura 6.4. Exemplo da estrutura do classificador, definido por um grafo direccionado acíclico.

Paralelamente ao processo de ligação de classes, é efectuado o cálculo da *Probabilidade de Evento Anormal* (P_{ea}) para cada nó. Esta probabilidade, que será utilizada posteriormente na identificação de eventos anormais, é obtida pela razão entre o número de trilhos de treino que passaram pelo nó e que originaram a violação de uma área restrita; e o número total de trilhos de treino que transitaram pelo mesmo nó, ou seja:

$$P_{ea} = \frac{\text{Número de trilhos com violação de área restrita}}{\text{Número total de trilhos}} \tag{6.15}$$

6.2.3. Mecanismo de Classificação de Comportamentos

Após a construção de um classificador com recurso a sequências de treino (trilhos), obtém-se uma estrutura estática que modela os comportamentos previamente observados. Cada novo trilho deverá ser avaliado de acordo com a informação existente no classificador. Para tal, é necessário verificar se os dados recolhidos são identificados como pertencentes a um dos nós de cada camada temporal correspondente. Este processo é efectuado pelo cálculo da distância de *Mahalanobis* às distribuições definidas pelo conjunto dos nós.

Considerando que numa determinada camada temporal existem S classes conectadas a um mesmo nó da camada temporal precedente, e que cada observação é definida por d atributos, então a distância de *Mahalanobis* entre uma observação multivariável $\mathbf{x}$ e uma distribuição $N_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, com $0<i<S$, é obtida a partir de um vector de médias $\boldsymbol{\mu}_i=\{\mu_1, \mu_2, \mu_3, \ldots, \mu_d\}$ e de uma matriz de covariância $\boldsymbol{\Sigma}_i$, de acordo com a seguinte equação:

$$D_i(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu}_i)^T \cdot \boldsymbol{\Sigma}_i^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu}_i)} \tag{6.16}$$

Como tal, a observação **x** é atribuída ao nó que contenha a classe que minimize a distância de *Mahalanobis*, desde que esta seja igual ou inferior a 1. Em caso de se verificar um valor de distância idêntico para duas ou mais classes, a observação **x** será atribuída à classe que apresentar maior densidade de probabilidade.

Quando um novo objecto entra em cena, este processo é realizado utilizando os nós da primeira camada temporal. Se não for satisfeita a condição de atribuição da amostra a um dos nós da camada inicial, considera-se que se está na presença de um evento não usual. Caso contrário, se a amostra é atribuída a um dos nós, procede-se à avaliação da *Probabilidade de Evento Anormal* (P_{ea}). Nos casos em que o valor da P_{ea} para o nó seleccionado seja superior a um limite predefinido, denominado por *Limite de Alarme* (A_l), então o classificador assinala a ocorrência de um evento anormal. No entanto, sempre que P_{ea} apresentar um valor igual ou inferior a A_l, avança-se no próximo espaço de tempo com o processo de associação, recorrendo aos "nós filhos" definidos na camada temporal seguinte.

Com a variação do parâmetro de *Limite de Alarme*, que toma valores do intervalo [0,1], é possível ajustar o classificador de modo a permitir a previsão da ocorrência de eventos de violação de espaços restritos. Com efeito, baixos valores de A_l possibilitam uma elevada antecipação na sinalização de eventos anormais. Por outro lado, quando A_l assume valores próximos do máximo, o classificador tende a perder a capacidade de previsão, realizando apenas a detecção de eventos de quebra de segurança.

A escolha do valor do parâmetro A_l deve ter em consideração as necessidades específicas de cada aplicação. Se por um lado em alguns ambientes se pretende sinalizar com enorme precocidade a possibilidade de violação de um espaço restrito, tolerando eventuais falsas detecções do classificador, em outros casos é dada preferência a uma reduzida taxa de erro em detrimento do nível de antecipação na detecção de eventos.

Através da análise de risco do local sob vigilância é possível tomar uma decisão sobre o valor a atribuir ao *Limite de Alarme*. Como tal, devem-se ponderar diversos factores como: a vulnerabilidade do local, o tipo e probabilidade de ocorrência de ameaças, bem como o impacto que estas possam causar. Cabe ao utilizador adoptar pela estratégia mais adequada a cada situação.

6.3. Classificador *DOG*

O classificador anteriormente proposto, i.e. *N-ary Trees*, apresenta um aspecto que se afigura como problemático na sua aplicação em ambiente real de vídeo-vigilância. O facto do modelo gerado permanecer estático após a fase de treino, ou seja, não possuir capacidade de adaptação, impede que sejam incorporadas no classificador alterações de padrões que se venham posteriormente a observar no ambiente sob monitorização. Em resposta a esta limitação, pretende-se desenvolver uma nova técnica que, respeitando a anterior abordagem, possibilite efectuar uma adaptação do modelo de classificação ao longo do tempo. Desenvolveu-se assim um novo classificador, denominado por *Dynamic Oriented Graph* (*DOG*), que substitui o método de construção de classes baseado em *Expectativa-Maximização* por uma outra técnica que será descrita ao longo desta secção.

O *DOG* foi idealizado com a finalidade de proporcionar um modelo de classificação adaptativo, i.e. exibir *plasticidade*, permitindo a geração de novas classes bem como possibilitar a sua reorganização caso os dados assim o determinem. Contudo, esta capacidade de incorporar novos padrões pode originar *instabilidade* nas estruturas das classes, tendo como efeito a perda de significado das classes resultantes. A esta contradição dá-se o nome de *dilema plasticidade/estabilidade* [Duda et al., 2001]. O classificador *DOG* supera este dilema pela integração de uma variante da *Teoria da Ressonância Adaptável* (*ART*[22]) *Gaussiana* [Williamson, 1996] com um método de fusão das classes que partilhem uma mesma região no hiperespaço definido pelos atributos.

6.3.1. Representação das Classes

Segundo a *Teoria da Ressonância Adaptável Gaussiana*, uma classe é definida através de uma distribuição gaussiana $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, apresentando um valor de média e de desvio padrão para cada dimensão. Esta informação é complementada por uma probabilidade *a priori* associada à respectiva classe, bem como pelo número de amostras que ela abarca.

Considerando M_c como sendo o número de amostras abrangidas pela classe W_c, de um total de C classes identificadas, então a probabilidade *a priori* de uma classe c, é definida por:

[22] Do inglês: *Adaptive Resonance Theory*.

$$P(W_c) = \frac{M_c}{\sum_{i=1}^{C} M_i} \tag{6.17}$$

As distribuições geradas a partir da aplicação da *ART Gaussiana* têm a particularidade de facultar um ajuste aos dados bastante satisfatório quando estes não apresentam uma correlação entre atributos, i.e. quando os atributos são independentes. Quando tal não se sucede, a sua influência pode ser minimizada através de uma análise prévia da correlação entre atributos. Com efeito, nos casos em que se verifique uma elevada correlação entre duas ou mais variáveis, é possível manter características discriminantes de cada classe utilizando apenas um dos atributos correlacionados e desprezando os restantes.

6.3.2. Selecção e Atribuição de uma Classe a um Vector de Entrada

À luz da *Teoria da Ressonância Adaptável Gaussiana* é possível a partilha de uma mesma região do espaço por várias distribuições gaussianas. Como tal, a associação de um vector de entrada**,** à classe que melhor o representa, é efectuada em duas fases. Inicialmente executa-se uma selecção da classe cuja distribuição constitua a procedência mais provável da amostra. Após a selecção, verifica-se se o critério de correspondência entre a amostra e a classe é satisfeito. Em caso afirmativo prossegue-se com a actualização dos parâmetros que definem a classe. Se tal não se verificar, uma nova classe é seleccionada.

A abordagem proposta para a construção do classificador *DOG* diverge neste aspecto da *Teoria da Ressonância Adaptável Gaussiana*. De modo a reduzir a complexidade deste processo, optou-se por uma solução que promove a fusão das classes cujas distribuições partilhem uma mesma região.

Considere-se **x** como o vector de entrada, de dimensão *d*, obtido pela média dos valores observados para cada atributo no período de tempo associado à camada temporal em análise. Considere-se ainda que na referida camada temporal se encontram definidas *C* classes identificadas por distribuições gaussianas $N_c(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, onde,

$$\Sigma_c = \prod_{i=1}^{d} \sigma_{c,i}^2 \tag{6.18}$$

Assim, a condição para a atribuição de uma classe a um vector de entrada **x**, é definida pelo seguinte critério:

$$\begin{cases} \text{atribuído} & , \text{ se } \left| x_{c,i} - \mu_{c,i} \right| \le \sigma_{c,i} \quad , \exists c \in C \wedge \forall i \in d \\ \text{não atribuído} & , \text{ caso contrário} \end{cases} \tag{6.19}$$

6.3.3. Método de Aprendizagem

Sempre que para um vector de entrada seja satisfeita a condição de atribuição a uma classe, esta deve ver actualizada a sua média, matriz de correlação e contador de amostras. No entanto, nos casos em que o vector de entrada não seja associado a qualquer classe (quer seja pela inexistência de classes na referida camada temporal, ou pelo não cumprimento do critério de atribuição), então uma nova classe é gerada.

A criação de uma classe implica a constituição de uma distribuição gaussiana cuja média é definida pelo vector de entrada (i.e. $\boldsymbol{\mu}=\mathbf{x}$), e por uma matriz de covariância baseada nos valores iniciais considerados para o desvio padrão em cada atributo. Como efeito, obtém-se um espalhamento isotrópico no espaço característico ao redor da primeira amostra da classe.

A técnica aqui proposta tem como finalidade a sua utilização num sistema de vídeo-vigilância real, onde a monitorização continuada pode proporcionar um elevado número de observações. Como tal, é indispensável que o processo de contagem de amostras seja dotado de mecanismos que permitam evitar problemas de implementação relacionados com a limitação de capacidade da variável que armazena o número de observações abarcadas por cada classe (i.e. problemas de *overflow*). Com este propósito desenvolveu-se uma técnica de contagem de amostras que possibilita não só evitar os problemas de limite de armazenamento da variável de contagem, como permite também implementar um método de "esquecimento" das classes que registem um número reduzido de observações.

Se M_c identificar a variável de contagem de amostras numa classe c, e $Mmax$ representar o valor máximo que a variável pode tomar, então quando a condição de atribuição é verificada para uma determinada classe, o seu contador de amostras é incrementado em uma unidade se o valor for inferior ao limite máximo, i.e $M_c=M_c+1$ se $M_c<Mmax$. Contudo, se se verificar que o contador tenha atingido o limite máximo ($M_c=Mmax$), então procede-se a uma redução, em uma unidade, nos contadores de todas as classes representadas pelo

classificador, i.e. M_n=M_n-1. Sempre que nesta operação um contador de amostras atingir o valor nulo (M_n=0), essa classe é removida do classificador.

Após a actualização do contador de amostras, o processo de aprendizagem da classe identificada prossegue com o ajuste dos valores de média e desvio padrão para cada atributo de modo a reflectir as propriedades da nova amostra.

$$\begin{aligned}&\text{Para cada} \quad i \in d \quad \text{fazer}\\ &\mu_{c,i} = \begin{cases}(1-\lambda)\cdot \mu_{c,i} + \lambda \cdot x_i & , \text{ se } \quad M_c > 1\\ x_i & , \text{ se } \quad M_c \le 1\end{cases}\\ &\sigma_{c,i} = \begin{cases}(1-\lambda)\cdot \sigma_{c,i} + \lambda \cdot \left|x_i - \mu_{c,i}\right| & , \text{ se } \quad M_c > 1\\ \sigma_{inicial} & , \text{ se } \quad M_c \le 1\end{cases}\end{aligned} \tag{6.20}$$

com, $0 \le \lambda \le 1$.

6.3.4. Fusão Entre Classes

Depois de concluído o processo de actualização das classes numa camada temporal, procede-se à verificação da existência de sobreposições entre as diversas distribuições representadas pelas classes aí definidas. Em caso afirmativo realiza-se uma fusão entre classes, duas a duas, de acordo com o especificado em (6.21).

$$\begin{aligned}&\text{Se, } \quad M_{c1} + M_{c2} < Mmax \quad , \text{ então}\\ &\quad M_c = M_{c1} + M_{c2}\\ &\text{Caso contrário,}\\ &\quad M_c = \frac{M_{c1} + M_{c2}}{2}\\ &\text{Para cada} \quad i \in d \quad \text{fazer}\\ &\quad \mu_{c,i} = \frac{M_{c1}}{M_c}\cdot \mu_{c1,i} + \frac{M_{c2}}{M_c}\cdot \mu_{c2,i}\\ &\quad \sigma_{c,i} = \frac{1}{2}\cdot \left[MAX\left(\mu_{c1,i} + \sigma_{c1,i}, \mu_{c2,i} + \sigma_{c2,i}\right) - MIN\left(\mu_{c1,i} - \sigma_{c1,i}, \mu_{c2,i} - \sigma_{c2,i}\right)\right]\end{aligned} \tag{6.21}$$

6.4. Geração de Dados Sintéticos

A avaliação dos dois classificadores (*N-ary Trees* e *DOG*), propostos no âmbito deste trabalho, deve ser realizada recorrendo a um mesmo conjunto de dados, de forma a permitir um estudo comparativo das duas técnicas. Idealmente a análise deveria ser efectuada com dados reais, obtidos num ambiente típico de vídeo-vigilância. No entanto optou-se por uma avaliação em duas etapas, em que numa primeira fase se utilizam dados sintéticos e, posteriormente, se realiza um teste com dados reais.

De acordo com o plano determinado para a avaliação, inicialmente os dois classificadores são analisados utilizando a um conjunto de dados sintéticos mas que possuam, tanto quanto o possível, características que se manifestam em ambientes reais (e.g. ruído, oscilações nas trajectórias, entre outros). Para o efeito desenvolveu-se uma aplicação semiautomática de geração de trilhos, denominada por *Observer – Track Generator* que, para além dessa funcionalidade, permite ainda a execução dos testes aos classificadores, bem como a visualização dos resultados dos mesmos. A Figura 6.5 apresenta uma vista desta aplicação.

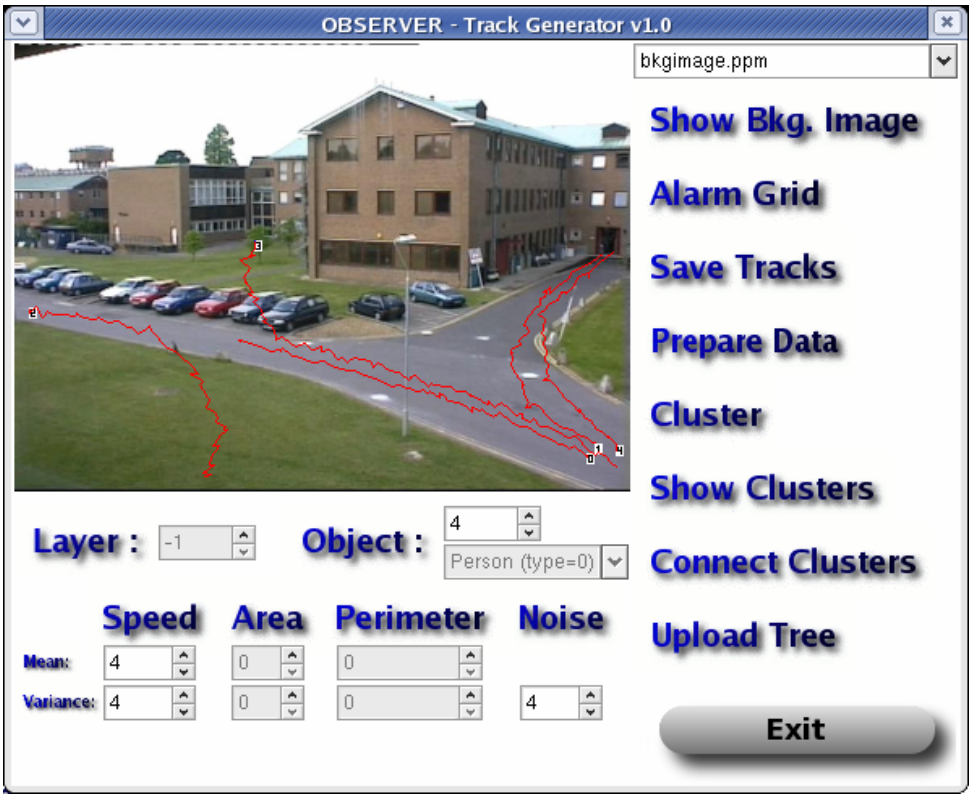


Figura 6.5. Aspecto da aplicação desenvolvida para a geração de trilhos assistida por utilizador.

A aplicação *Observer – Track Generator* foi concebida de modo a facultar um meio de geração de trajectórias afectadas por ruído gaussiano nas coordenadas do centro-de-gravidade do objecto, na sua velocidade, área e perímetro. Assim, pela introdução das coordenadas de início e fim de percurso, bem como da identificação dos pontos de mudança de direcção, é construída uma trajectória completa, semelhante a uma trajectória típica de ambiente real. Veja-se como exemplo a Figura 6.6 (a), onde é representado um trilho gerado por esta aplicação, obtido pela selecção de apenas quatro coordenadas da imagem (pelo utilizador), e que produziu automaticamente uma sequência de 92 pontos.

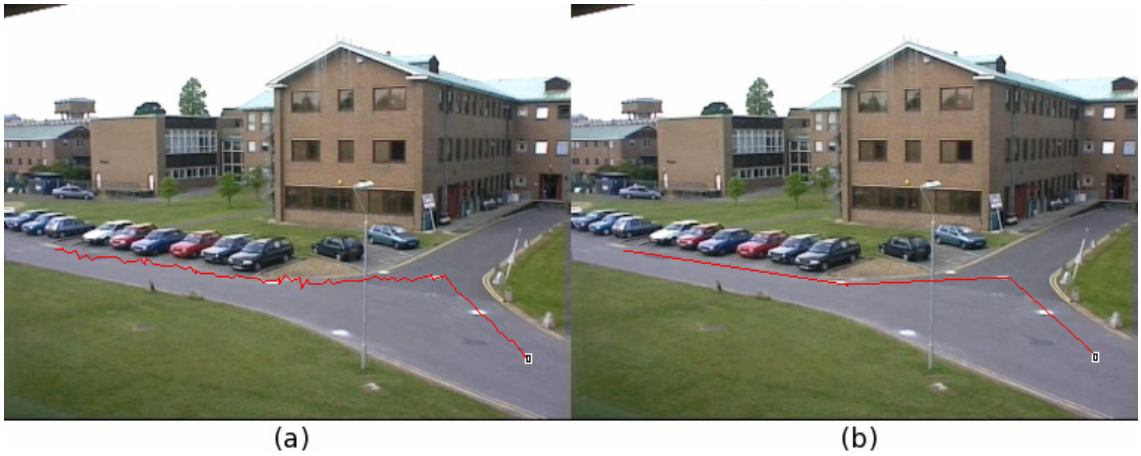


Figura 6.6. (a) Trajectória gerada pela aplicação *Observer – Track Generator* com selecção de apenas quatro coordenadas na imagem; (b) A mesma trajectória sem a introdução de ruído.

A aplicação desenvolvida permite ainda a selecção de áreas consideradas restritas a cada uma das classes de objectos (i.e. pessoas, grupos de pessoas, e veículos). Para tal, a imagem que representa o ambiente a monitorizar é dividida numa grelha, utilizada pelo operador para definir as regiões restritas. Como resultado, é gerado um mapa contendo as áreas sensíveis do ponto de vista da segurança. Tendo por base esta informação, é então possível rotular as trajectórias posteriormente geradas como normais ou anormais.

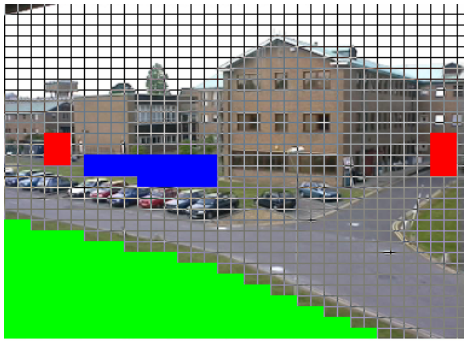


Figura 6.7. Exemplo de grelha de selecção de áreas restritas a pessoas, grupos e veículos.

Através do uso do *Observer – Track Generator* foi construído um conjunto de dados para avaliação dos classificadores de comportamentos, baseado no ambiente utilizado pela PETS 2001, composto por 1000 trilhos de diferentes comprimentos. Para o efeito, foram também definidas duas áreas restritas a pessoas, como assinalado na Figura 6.8 (b). Os trilhos gerados representam 16 trajectos distintos de um mesmo tipo de objecto (i.e. pessoa), em que 20% dos trilhos violam uma das áreas restritas.

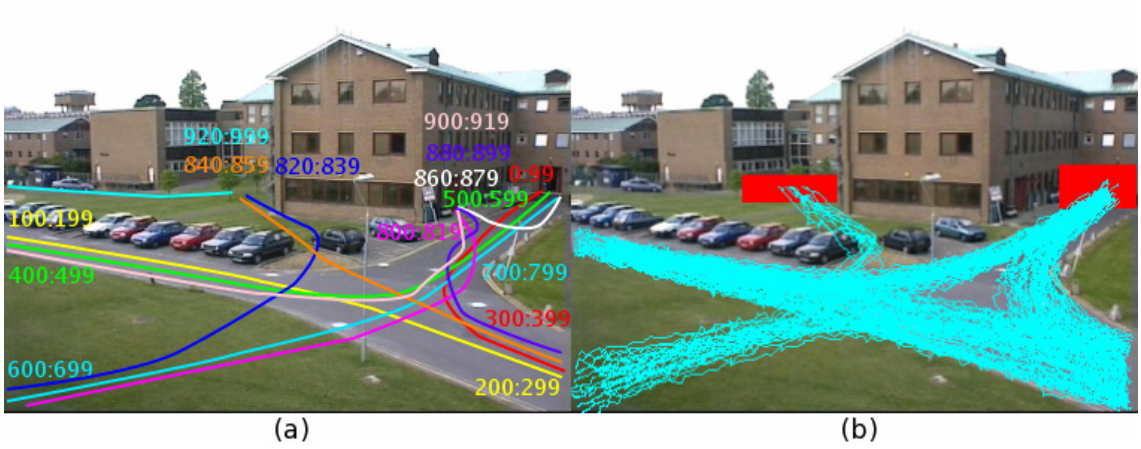


Figura 6.8. (a) Representação das 16 trajectórias definidas por 1000 trilhos; (b) Alguns dos trilhos sintéticos que compõem o conjunto de dados de teste.

6.5. Teste dos Classificadores Sobre Dados Sintéticos

Para a análise dos classificadores *N-ary Trees* e *DOG* adoptou-se um esquema de *validação cruzada* do tipo *10-fold cross validation* [Kohavi, 1995]. O classificador *DOG* foi avaliado utilizando a cada etapa do processo 9/10 do conjunto de dados para a aprendizagem dinâmica do modelo e os restantes 1/10 para teste do classificador com inibição da capacidade de adaptação dinâmica. Com o *N-ary Trees* o processo foi similar, realizando uma fase de treino com 9/10 dos dados e aplicando os dados remanescentes para o teste do modelo.

De modo a aumentar o grau de confiança dos resultados da avaliação, realizaram-se 30 diferentes execuções de *10-fold cross validation* para cada classificador. Assim, o teste de um classificador implica a simulação de 30 processos de validação, o que perfaz uma soma total de 3 milhões de trilhos a processar, i.e. 1000 trilhos por 30 execuções de *validação cruzada*, onde são testados 100 valores distintos para o parâmetro de *Limite de Alarme* (A_l).

Os testes foram executados sob as mesmas condições, num computador baseado em processador de 32 *bits* a 3.0GHz, com 512MB de memória RAM, e operando um sistema *GNU/Linux*. Utilizando uma plataforma de *hardware* comum para a avaliação, é possível comparar o desempenho em termos de recursos temporais inerentes a cada abordagem de classificação.

Para além do tempo consumido por cada classificador, pretende-se ainda aferir os seus níveis de eficácia na previsão de comportamentos. Como tal, foram calculadas as curvas *ROC*[23] [Fawcett, 2003] e as curvas de *Taxa de Antecipação*. As curvas ROC foram calculadas pela construção de uma *Matriz de Confusão* para cada valor de *Limite de Alarme* (A_l), com A_l adoptando valores compreendidos entre [0%, 100%], fraccionado em intervalos de 1%. A *Taxa de Antecipação* é obtida pelo cálculo da razão entre a distância de antecipação (i.e. distância que vai desde o ponto onde foi efectuada a previsão de evento anormal até ao ponto onde ocorre a violação efectiva de um espaço restrito) e a distância total desde o início do trilho até à sua entrada numa área restrita. A *Taxa de Antecipação* é calculada para diferentes valores de A_l, de modo similar ao que acontece com as curvas *ROC*.

Para cada processo de *validação cruzada*, efectuou-se a medição do tempo necessário à sua execução, bem como do número de *Verdadeiros Positivos*, *Falsos Positivos*, *Verdadeiros Negativos* e *Falsos Negativos*. Considerou-se ainda uma *Distribuição t-student* para o cálculo dos intervalos de confiança dos *Verdadeiros* e *Falsos Positivos* [Flexer, 1996].

A realização total do teste para o classificador *DOG* requereu somente 1 minuto e 29 segundos, enquanto que a execução da avaliação do *N-ary Trees* exigiu elevados recursos computacionais, acumulando um tempo total de execução de aproximadamente 190 horas. Esta enorme disparidade de tempo pode ser explicada pelo considerável custo computacional exigido pelo classificador *N-ary Trees* durante a fase de treino.

Na Figura 6.9 é apresentada a curva *ROC* para o classificador *N-ary Trees* onde se exibe, para cada valor de A_l, as variações esperadas para um intervalo de confiança de 95%, enquanto que na Figura 6.10 se mostram os resultados obtidos para o classificador *DOG*. Note-se que a linha tracejada (a vermelho) representa a curva *ROC* de um classificador aleatório.

---

[23] *Receiver Operating Characteristic.*

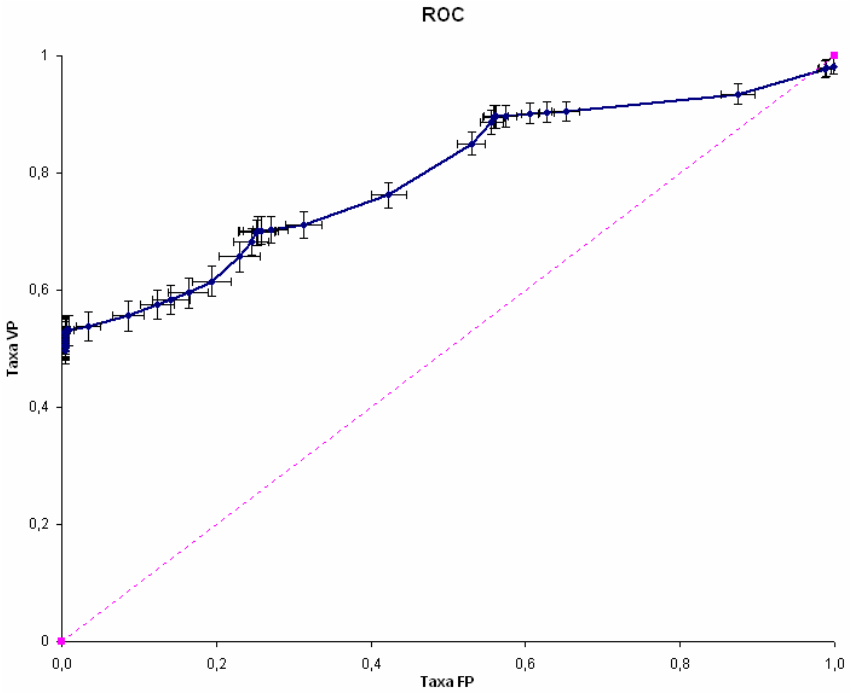


Figura 6.9. Curva *ROC* para o classificador *N-ary Trees*, com apresentação das variações esperadas para um intervalo de confiança de 95%.

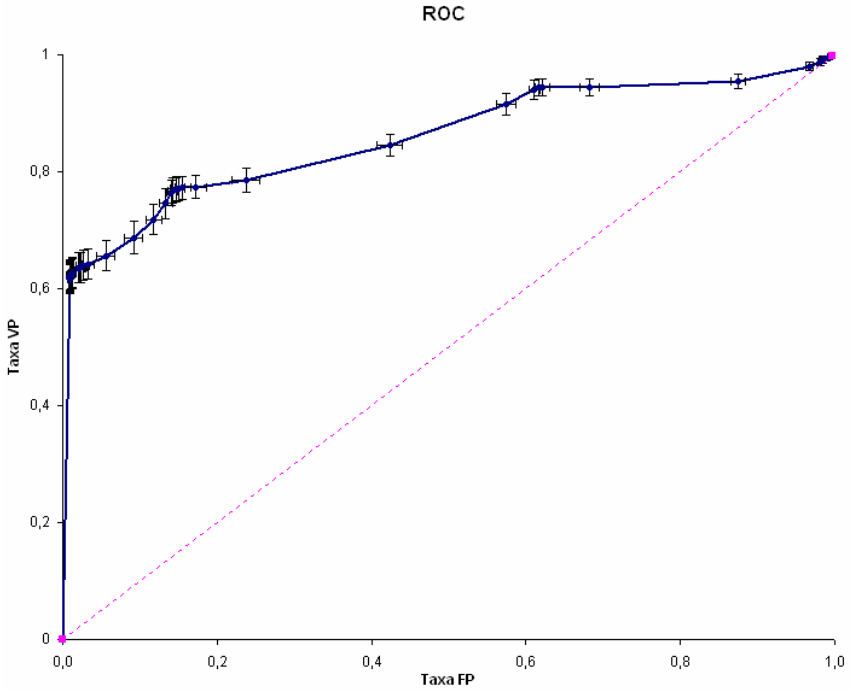


Figura 6.10. Curva *ROC* para o classificador *DOG*, com apresentação das variações esperadas para um intervalo de confiança de 95%.

Uma outra métrica de avaliação do desempenho de classificadores utilizada neste estudo foi a *AUC* (do inglês: *Area Under the Curve*) da curva *ROC*. Esta métrica permite uma avaliação comparativa de classificadores que processam um conjunto de dados comum, sendo que o classificador ideal apresenta uma *AUC* de 100%. Com efeito, nos testes efectuados, observou-se um valor de *AUC* na ordem dos 86% para o classificador *DOG*, em contraste com o valor de 79% obtido no teste do classificador *N-ary Trees*.

Na Figura 6.11 é possível observar, para o domínio de valores possíveis de A_t, o comportamento dos dois classificadores propostos. Como se pode verificar, a *Taxa de Verdadeiros Positivos* decai à medida que A_t aumenta. Constata-se ainda que a *Taxa de Antecipação* diminui para valores mais elevados do *Limite de Alarme*. Tal comportamento é comum a ambos os classificadores.

Suportando-se na informação fornecida pela Figura 6.11, é facilitada a tomada de decisão em relação ao valor ideal para o parâmetro de *Limite de Alarme*. A combinação da informação da curva *ROC* com a curva da *Taxa de Antecipação* permite definir um valor óptimo, baseado no compromisso entre a *Taxa de Falsos Positivos*, *Taxa de Verdadeiros Positivos* e nível de antecipação requerido.

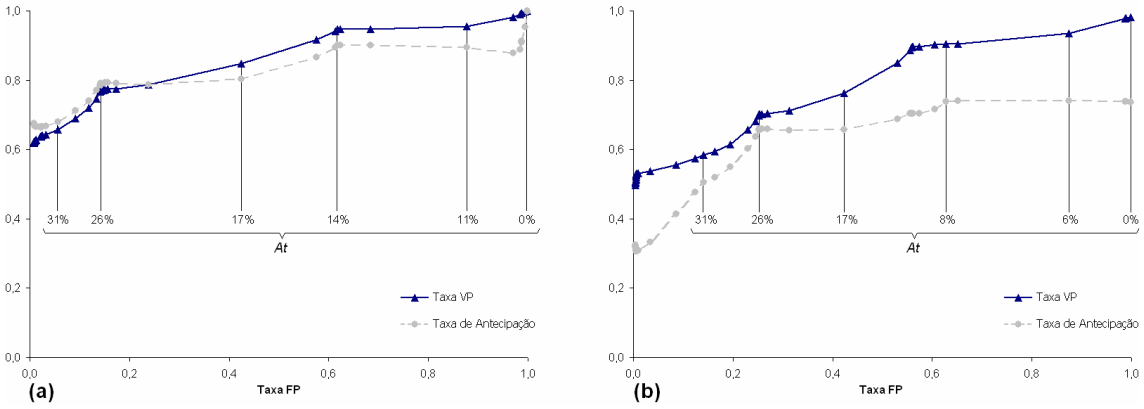


Figura 6.11. Curva *ROC* sobreposta pela curva de *Taxa de Antecipação* para: (a) classificador *DOG* e (b) classificador *N-ary Trees*. (Eixo dos y identifica simultaneamente a Taxa de *VP* e Taxa de Antecipação).

6.6. Protótipo do Sistema

De acordo com o propósito deste trabalho, pretendia-se construir um protótipo do sistema de detecção e previsão de comportamentos anormais, a implementar numa câmara inteligente. Tal câmara deve apresentar algumas características específicas, nomeadamente:

- Sensor CCD ou CMOS a cores com resolução mínima de 640x480, codificação a 24bits por ponto, e com uma taxa de aquisição de 25 imagens por segundo (fps);
- Varrimento progressivo (em inglês "*progressive scan*");
- Sistema operativo *GNU/LINUX* com possibilidade de adição de aplicações por parte do utilizador;
- Processador com 32bits a 3.0GHz ou equivalente;
- Interface de rede 100Mb/s ou superior.

Pese embora o rumo da investigação em análise de vídeo, bem como da tendência da indústria da vídeo-vigilância, apontarem para a necessidade de distribuição do processamento para as câmaras de vídeo, não existe actualmente uma plataforma de *hardware* que permita implementar um sistema de tal complexidade computacional. Pesquisando as soluções disponibilizadas actualmente pelos fabricantes de equipamentos de vídeo-vigilância, a nível mundial, verifica-se a não existência de uma plataforma que suporte a tecnologia desenvolvida no âmbito deste doutoramento.

Recentemente têm sido apresentadas algumas câmaras de vídeo-vigilância IP, com aquisição de imagem por varrimento progressivo de média e elevada resolução. Exemplos destes equipamentos são os modelos 210, 210A e 211 da Axis[24] e as séries M12 e M22 da Mobotix[25]. Embora estes modelos possam constituir a solução mais avançada actualmente disponível, estes não se adequam às necessidades de processamento dos algoritmos propostos e constituem soluções fechadas, impossibilitando a incorporação na própria câmara de qualquer programação realizada pelo utilizador.

A solução passa então por utilizar um outro tipo de câmaras de vídeo, empregues em visão por computador para aplicações industriais de teste e medida. Tratam-se de câmaras digitais

[24] URL: www.axis.com

[25] URL: www.mobotix.com

de alta performance, cuja instalação, manuseamento e programação requerem um elevado conhecimento técnico. Outras limitações advêm do facto de que este tipo de câmaras serem projectadas para funcionamento em ambientes laboratoriais controlados, e apresentarem ainda um custo extremamente elevado, quando comparadas com os equipamentos utilizados para vídeo-vigilância.

Obviamente não se espera que estas câmaras constituam a solução ideal para as futuras câmaras inteligentes, no entanto, são o único meio actualmente disponível para implementação e teste da abordagem proposta para a detecção e previsão de comportamentos anormais.

Após uma exaustiva pesquisa de equipamentos de aquisição de vídeo industrial, adequados ao caso em questão, obtiveram-se os seguintes equipamentos:

- Fabricante: Sony
  Modelo: XCI-V3
  Sensor: CCD, Monocromático, 60fps (VGA)
  Varrimento: Progressivo
  Sistema Operativo: Windows XPe ou Linux
  Processador: x86 AMD Geode GX5333 400MHz
  Interface de Rede: 10 / 100Mb/s
- Fabricante: Tattile
  Modelo: XP TAG
  Sensor: CCD, Cores, 60fps (VGA) e 15fps (1600x1200)
  Varrimento: Progressivo
  Sistema Operativo: Windows XPe ou Linux
  Processador: Dual Core DSP
  Interface de Rede: 10 / 100Mb/s
- Fabricante: Matrix Vision
  Modelo: mvBlueCOUGAR-P
  Sensor: CCD, Monocromático ou Cores, 640x480 a 1280x1024
  Varrimento: Progressivo
  Sistema Operativo: Windows XPe ou Linux
  Processador: PowerPC a 400MHz
  Interface de Rede: 10 / 100 / 1000Mb/s

- Fabricante: Basler
  Modelo: EXA640-60C
  Sensor: CMOS, Cores, 60fps (656x490)
  Varrimento: Progressivo
  Sistema Operativo: Linux (Kernel 2.6)
  Processador: PMC-Sierra RM9000 (64bits) a 1GHz
  Interface de Rede: 10 / 100 / 1000Mb/s

A escolha da plataforma de implementação da câmara inteligente recaiu sobre o modelo EXA640-60C da Basler, uma vez que se trata do equipamento que melhor satisfaz os requisitos apresentados. Também teve especial importância nesta escolha a elevada capacidade de processamento do seu processador PMC-Sierra RM9000.

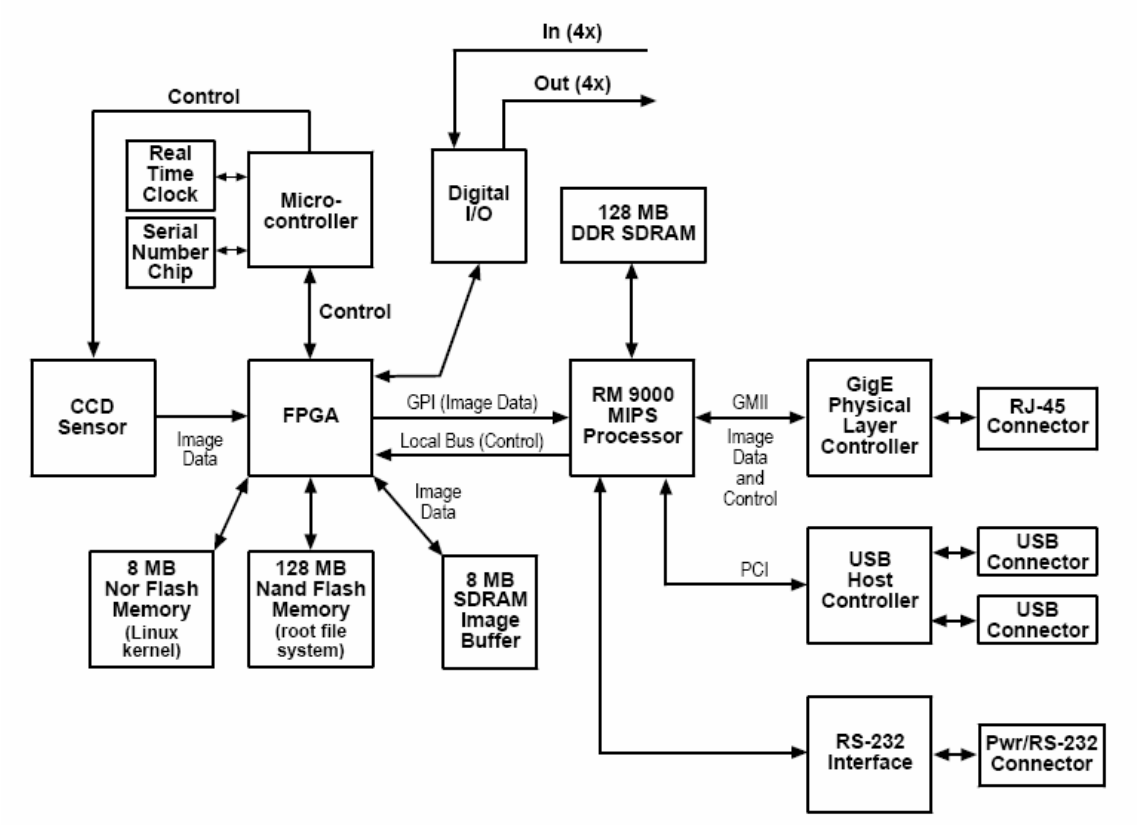


Figura 6.12. Diagrama de blocos da câmara EXA640-60C da Basler.

Efectuada a selecção da plataforma de *hardware*, procedeu-se então à compilação do código fonte para um processador 64bits-MIPS. Após a realização de um teste preliminar, verificou-se a necessidade de efectuar a adaptação de algumas funções do código fonte, devido a restrições impostas pela arquitectura da câmara de vídeo, nomeadamente, à frequente ocorrência de *stack overflow* na execução de funções que apresentam recursividade, obrigando assim à reescrita de todas essas funções.

6.6.1. Modelo Computacional Cliente – Servidor

Com a aparição das câmaras de vídeo IP, cuja transmissão da imagem é efectuada por rede através da implementação do protocolo *TCP/IP*, assistiu-se à adopção por parte da indústria da vídeo-vigilância do modelo computacional Cliente – Servidor. Neste novo modelo, a interligação entre as câmaras e o equipamento concentrador de vídeo (e.g. *DVR*, *NVR* ou *HVR*[26]) é assegurada por uma normal infra-estrutura de rede de computadores. Como se uma vulgar rede informática se tratasse, as imagens vídeo são codificadas e posteriormente transmitidas através de cablagem *UTP*, fibra óptica ou tecnologia *Wi-Fi*.

Nesta nova abordagem as câmaras de vídeo são identificadas como "servidores", aguardando a requisição de um determinado serviço por parte do dispositivo "cliente". As câmaras recebem as requisições de um serviço, avaliam a sua legitimidade, processam as requisições e retornam o resultado ao cliente. A Figura 6.13 ilustra este modelo, com a câmara tendo o papel de "servidor" aos pedidos efectuados pelo equipamento de recepção de imagens ("cliente").

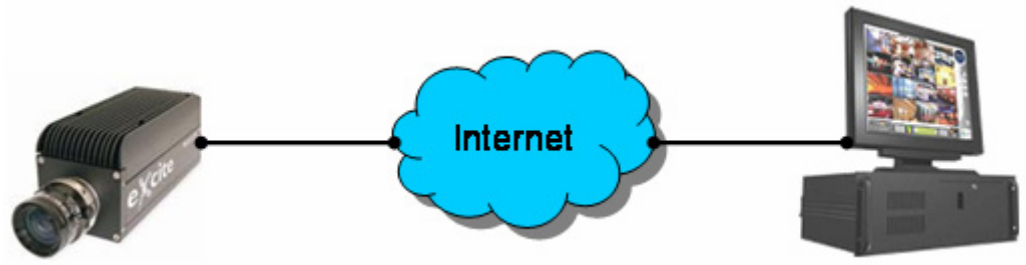


Figura 6.13. Modelo computacional Cliente – Servidor para um sistema *CCTV* de quarta geração.

[26] *DVR*, *NVR* e *HVR* são respectivamente siglas de *Digital*, *Network* e *Hybrid Video Recorder*.

A câmara inteligente proposta no âmbito deste trabalho implementa igualmente o modelo computacional Cliente – Servidor. Note-se no entanto que as requisições realizadas pelos equipamentos "cliente" não se limitam a pedidos de envio de imagens vídeo. Uma câmara inteligente assegura um leque mais alargado de serviços, fornecendo em simultâneo um conjunto de meta-informação descritiva da cena observada.

Para garantir a comunicação entre cliente e servidor é necessário definir um protocolo de comunicação que assegure a implementação dos serviços desejados, nomeadamente o envio de uma sequência de vídeo acompanhada da meta-informação descritiva.

O *MPEG-7*, também conhecido por Interface para Descrição de Conteúdo Multimédia, é um padrão que efectua a modelação em *XML* de meta-informação relativa a um conteúdo multimédia. O *MPEG-7* é padrão independente, isto é, não obriga à utilização de padrões *MPEG* de compressão de áudio/vídeo, podendo mesmo ser utilizado na descrição de conteúdo de vídeo analógico.

Embora o padrão *MPEG-7* tenha sido desenvolvido como objectivo de fornecer um método rápido e eficaz para a procura, filtragem e identificação de conteúdos, permitindo ainda realizar a descrição de modelos, estruturas e características de baixo-nível, este poderia ser utilizado para a descrição dos eventos observados pelo sistema de previsão e identificação de comportamentos anormais. Todavia, e apesar de reconhecer que a sua utilização seria apropriada numa solução comercial de uma câmara inteligente, tal implementação não se justifica nesta fase de prova de conceito do sistema.

Em alternativa, optou-se assim pela definição de um protocolo básico, que assegurasse as seguintes funcionalidades:

- configuração de ganho, brilho;
- configuração de número de imagens capturadas por segundo;
- leitura e envio de configurações do sistema;
- pedido de imagens vídeo;
- pedido de iniciar e finalizar detecção de comportamentos;
- comutação entre modo de aprendizagem, e modo de detecção; e
- pedido de leitura da estrutura do classificador.

Definiu-se então um formato para a estrutura de troca de mensagens entre a aplicação cliente e servidor. Assim, a aplicação cliente efectua a requisição de um serviço, enviando uma mensagem (encapsulada em TCP/IP) contendo dois campos:

<ID_Pedido/Resposta> <Dados>

Alguns pedidos implicam o envio de uma mensagem de resposta por parte do servidor onde, neste caso, o campo "Dados" pode conter uma imagem capturada pela câmara de vídeo, as configurações em uso pelo sistema, ou a estrutura do classificador *DOG*.

A seguinte tabela apresenta os pedidos da aplicação cliente implementados neste protocolo.

Tabela 6.1. Tabela de pedidos da aplicação cliente.

PEDIDO	MENSAGEM	RESPOSTA
Ganho	IDMSG01 <[0,255]>	-
Brilho	IDMSG02 <[0,1023]>	-
Imagens por Segundo	IDMSG03 <[1,25]>	-
Capturar Vídeo	IDMSG04 X	IDRSP01 <IMAGEM>
Iniciar Observer	IDMSG05 X	IDRSP01 <IMAGEM>
Reiniciar Observer	IDMSG06 X	IDRSP01 <IMAGEM>
Receber Configurações	IDMSG07 X	IDRSP02 <CONFIG>
Enviar Configurações	IDMSG08 <CONFIG>	-
Receber DOG	IDMSG09 X	IDRSP03 <DOG>

Como se pode verificar pela tabela, o parâmetro de "ganho" toma valores compreendidos entre 0 e 255, enquanto que o brilho aceita valores entre 0 e 1023. O número de imagens capturadas por segundo é também configurável entre 1 a 25 imagens. Na mensagem de envio de configurações, o campo "CONFIG" tem o seguinte formato:

THRESHOLD PERCENT TH TS ALFASM BETASM ALFALM BETALM AREA MINVEHICLEAREA OUTPUTIMAGE

Os dois primeiros campos, i.e. "THRESHOLD" e "PERCENT", referem-se ao processo de segmentação de movimento. O primeiro campo, que toma valores entre 0 e 255, diz respeito ao nível de sensibilidade para a sinalização de movimento num determinado ponto

da imagem, de acordo com o especificado pela equação (4.11). Em "PERCENT" ajusta-se o parâmetro de detecção de deslocação da câmara de vídeo, onde se define a percentagem de movimento a partir da qual se considera ter ocorrido uma movimentação da câmara de vídeo ou alteração do ponto de focagem.

Os campos "TH" e "TS" estão respectivamente associados à configuração da variação das componentes de matriz e de saturação, no espaço de cor *HSV*, para a detecção de sombras e brilhos. Ambos os parâmetros tomam valores compreendidos pelo intervalo de 0 a 255, segundo as equações (4.4) e (4.5).

Ainda relacionados com a detecção de sombras encontram-se os parâmetros "ALFASM" e "BETASM", que definem os limites de variação da componente de valor do espaço *HSV*, como representado pela equação (4.6). De forma análoga procedesse à atribuição dos parâmetros de detecção de brilhos, i.e. "ALFALM" e "BETALM", de acordo com a equação (4.7).

O campo "AREA" define o número de pontos mínimo, a partir do qual uma região segmentada é considerada como válida para os processos de seguimento e classificação de objectos em movimento. O parâmetro "MINVEHICLEAREA" define a área mínima, em pontos, para um veículo.

O campo "OUTPUTIMAGE" permite seleccionar o tipo de imagem a receber. Ajustando este valor, é possível receber imagem em: formato *RGB*, *HSV* ou tons de cinzento; *RGB* ou *HSV* de plano de fundo; imagem binária da Máscara Primária de Movimento; imagem binária da Máscara de Movimento; imagem binária da Máscara de Sombras; imagem binária da Máscara de Brilhos; imagem binária de contornos; imagem *RGB* com modelos de aparência; e imagem *RGB* com identificação e classificação de comportamentos.

Para se poder observar as imagens provenientes da câmara em tempo-real, e de modo a facultar um fácil e eficaz controlo sobre as funcionalidades da câmara inteligente, desenvolveu-se um interface gráfico do utilizador, aqui apresentado na Figura 6.14. Esta interface permite aceder a todos os serviços disponibilizados, facultando ainda funções como a selecção de áreas restritas para cada tipo de objectos (i.e. pessoas, grupos e veículos).

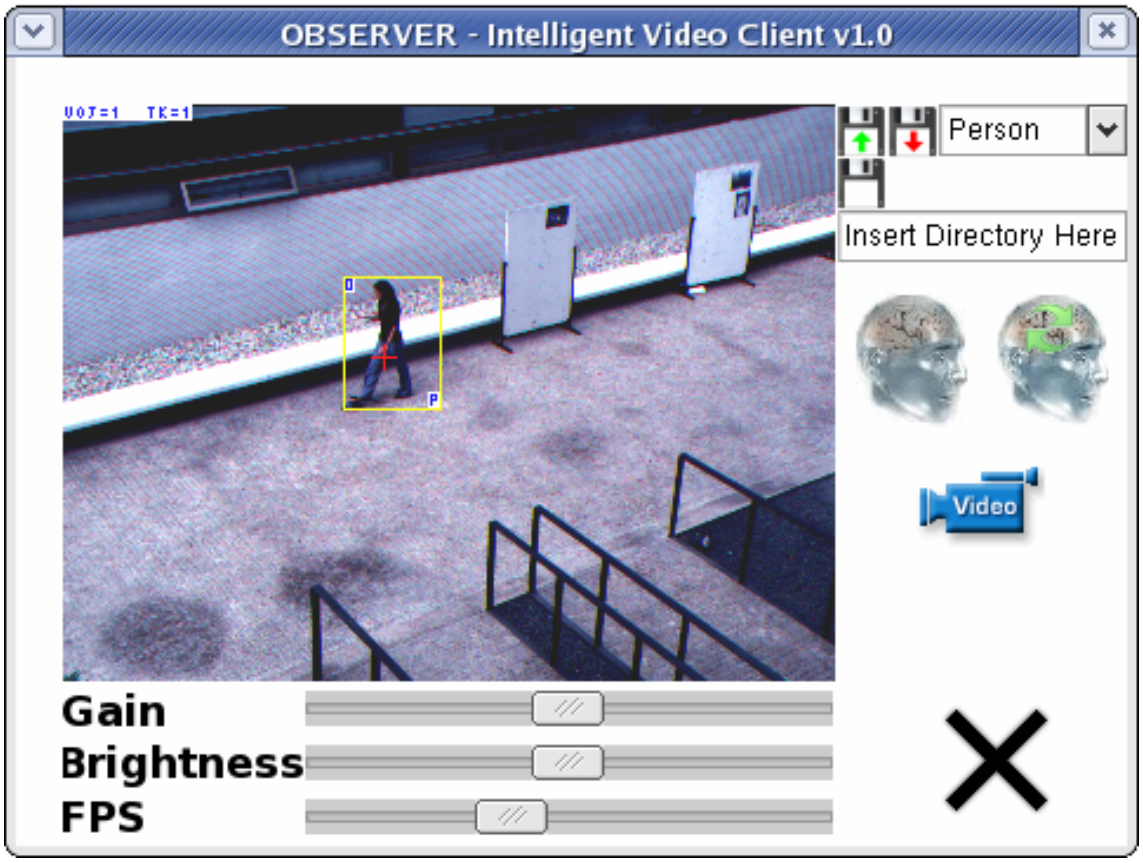


Figura 6.14. Interface gráfico do utilizador para a aplicação cliente.

6.7. Teste do Sistema em Ambiente Real

Num local sob vigilância, como por exemplo um balcão de uma instituição financeira ou uma sala de um museu, os eventos de quebra efectiva de segurança (**eventos anormais**) são extremamente raros, podendo mesmo nunca serem observados durante um longo período de tempo (meses ou anos). Como exemplo veja-se os casos dos bancos e instituições financeiras que, segundo a Associação Portuguesa de Bancos, contavam-se 5977 balcões em Portugal no ano de 2007 [APB, 2008]. Deste universo, registaram-se apenas 153 assaltos. Se se examinarem as ocorrências de assaltos em terminais *ATM*, verifica-se que do universo de 12510 terminais existentes em Portugal, foram realizados 64 roubos, ou seja, uma percentagem semelhante à verificada nos balcões de 0,005% (assaltos/balcão).

Como se pode constatar, a instalação do protótipo desenvolvido num local público, com relevância do ponto de vista da manutenção da segurança, é incompatível com as limitações

impostas pela calendarização de um trabalho de doutoramento. Existe uma reduzida probabilidade de ocorrência de assaltos o que poderia levar anos até se poder observar um **evento anormal**.

Uma abordagem alternativa poderia passar pela utilização de actores que, executando actos previamente definidos, protagonizariam um assalto ou um outro acto de quebra de segurança, como a violação de um espaço restrito. Contudo, situações encenadas podem não corresponder ao comportamento verdadeiro de um criminoso ou infractor.

Para resolver tal problema, optou-se por uma solução que passa pela definição de uma determinada acção observada com pouca frequência, como sendo de facto um evento anormal. Tal solução é conseguida pela selecção de uma área restrita (a pessoas) onde se verificaram previamente um reduzido número de movimentações. Deste modo, os eventos de quebra de segurança são observados com maior assiduidade, não sendo viciados de nenhuma forma.

Ao contrário dos casos de quebra de segurança, os **eventos não usuais** são mais frequentemente observados. Situações como pessoas correndo pelo átrio de um edifício, paradas por um período anormalmente longo ou descrevendo uma trajectória nunca antes observada, apesar de não constituírem eventos anormais, verificam-se com maior incidência.

Os **eventos não usuais**, que teriam uma importância vital na tarefa de alertar o vigilante para uma possível ameaça, não serão no entanto contabilizados no teste real. Tal facto justifica-se pelo objectivo principal deste trabalho se centrar na previsão e antecipação de uma **violação de um espaço restrito**.

Com este teste espera-se medir a evolução do número de eventos anormais ao longo do tempo, verificando o tempo necessário para a estabilização do sistema. Pretende-se ainda visualizar a dispersão dos objectos pelo espaço monitorizado.

6.7.1. Cenário de Avaliação

Pretendia-se efectuar o teste real num espaço amplo, com um elevado fluxo de movimento, onde fosse observável um grande número de trajectórias e comportamentos. Com este propósito, identificou-se o átrio do edifício da Escola de Engenharia da Universidade do Minho, Campus de Azurém, como um local favorável à realização deste teste. Assim, foi

assinalado o ponto onde seria fixada a câmara inteligente e tomadas as medidas necessárias para a instalação de alimentação eléctrica e infra-estrutura de rede. Procedeu-se ainda à submissão de pedidos de autorização à Comissão Nacional de Protecção de Dados (*CNPD*) e aos Serviços Técnicos da Universidade do Minho (*STUM*).

Contudo, e apesar do edifício já estar equipado com sistema *CCTV*, a instalação da câmara inteligente no exterior do edifício não foi autorizada. A escolha do cenário para a avaliação do sistema ficou então circunscrita ao espaço observável a partir do laboratório onde o trabalho foi desenvolvido. Esta restrição foi imposta pelos *STUM*, por considerar que este equipamento poria em causa a segurança de pessoas e bens.

De acordo com as restrições impostas, seleccionou-se uma área (do átrio do edifício da Escola de Engenharia) que torna possível observar a entrada e saída de pessoas do edifício. Esta área tem especial relevância por ser utilizada pelos alunos como espaço de convívio, onde é possível observar diferentes tipos de comportamentos e trajectórias. A Figura 6.15 apresenta o espaço escolhido como cenário de avaliação do sistema.

Figura 6.15. (a) Imagem do local sob monitorização para o teste real do sistema. (b) Fotografia da instalação da câmara inteligente.

A câmara de vídeo inteligente é colocada numa posição a 5,5 metros da base do local sob monitorização, e direccionada segundo uma inclinação de cerca de 70 graus (a partir da normal com a superfície alvo) de modo a observar uma área útil de aproximadamente 80 metros quadrados, como se mostra na Figura 6.15 (a). Com tal disposição, os objectos em movimento pela zona sob vigilância distam no mínimo em 7 metros, sendo que a distância máxima à câmara de vídeo se situa por volta do 18 metros.

6.7.2. Escolha de Parâmetros do Sistema

A selecção dos parâmetros do sistema deve ser realizada de forma criteriosa, com o objectivo de assegurar o melhor resultado possível na segmentação e seguimento dos objectos em movimento.

Alguns parâmetros são considerados constantes, e definidos como valores fixos no próprio código fonte do sistema. São o caso de:

- $\alpha = 0.9$, na equação (4.20);
- $T = 12$, na equação (4.17);
- $\alpha = 0.9$, nas equações (5.17) e (5.18).

Apesar da maioria dos parâmetros possuírem a característica de constância de um valor óptimo para um elevado número de cenários de vigilância, alguns valores podem ser ajustados de modo a optimizar o desempenho do sistema em condições específicas (e.g. condições de iluminação com extrema variabilidade da intensidade luminosa). Como tal, os valores aqui atribuídos devem ser entendidos como uma atribuição de parâmetros em situações "padrão", podendo ser adaptados de acordo com cada caso.

Os parâmetros cuja configuração é permitida são os seguintes:

- $T = 16$, na equação (4.11);
- $\tau_H = 30$, na equação (4.4);
- $\tau_S = 30$, na equação (4.5);
- $\alpha = 0.75$ e $\beta = 0.97$, na equação (4.6);
- $\alpha = 1.03$ e $\beta = 1.53$, na equação (4.7);
- AREA = 400;
- MINVEHICLEAREA = 2000.

6.7.3. Resultados Experimentais

Num ambiente real, é comum observar-se uma reacção causa-efeito, na evolução da cadeia dos acontecimentos. Por outras palavras, uma acção protagonizada por um dos objectos em movimento pode causar uma reacção, por interacção directa ou indirecta com outros objectos que partilhem o mesmo espaço. Como tal, a avaliação real do sistema deve ter em

consideração a ordem pela qual os objectos aparecem em cena. Este factor tem influência na construção do classificador, podendo ter preponderância nos resultados por ele produzidos.

Em tais circunstâncias, o esquema utilizado para a análise dos classificadores *N-ary Trees* e *DOG*, i.e. o esquema de *validação cruzada* do tipo *10-fold cross validation* [Kohavi, 1995], não se apresenta como o mais aconselhado para a avaliação do teste real. Isto porque este método realiza uma reorganização dos dados (através de uma mistura aleatória) antes da sua estruturação em grupos.

Optou-se então por um método denominado por *re-treino incremental*, do inglês *Incremental Retraining*, que encontra aplicação na avaliação de sistemas cuja ordem temporal deva ser mantida. Um exemplo deste método pode ser encontrado em [Metsis et al., 2006] na avaliação de um algoritmo *Anti-Spam*.

O *re-treino incremental* consiste na segmentação de uma sequência ordenada de dados, em lotes de tamanho fixo (com *k* trilhos adjacentes). O primeiro lote encerra os *k* trilhos mais antigos, sendo que o segundo engloba os *k* trilhos observados imediatamente após os trilhos do primeiro lote, e assim sucessivamente.

O procedimento de avaliação com *re-treino incremental* é implementado de acordo com o esquema representado na Figura 6.16. Neste método, os dados adquiridos são repartidos em *n* lotes, cada um contendo um igual número *k* de dados adjacentes. A ordem pela qual os dados são adquiridos é preservada.

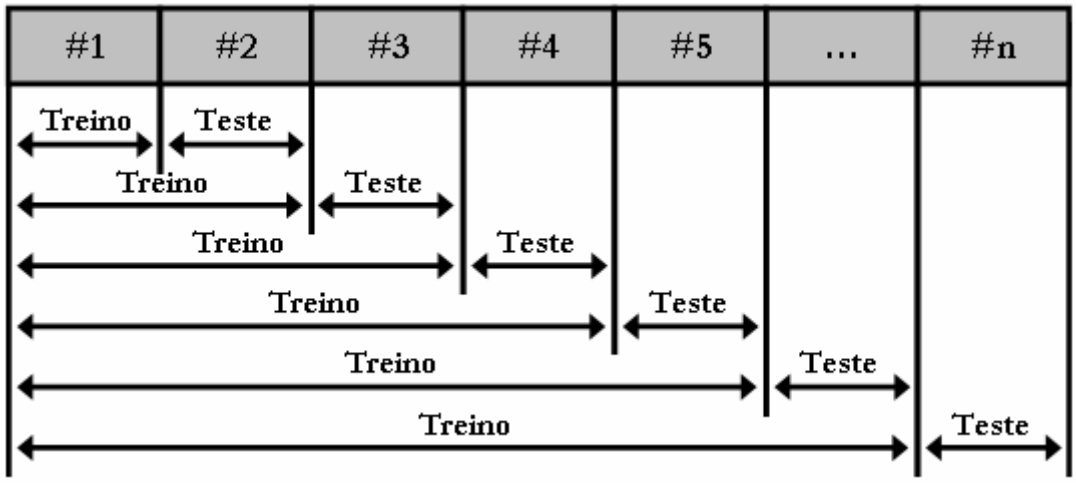


Figura 6.16. Esquema de utilização dos grupos de dados (lotes) utilizados no *re-treino incremental*.

O procedimento de avaliação do classificador consiste em realizar *n*-1 sequências de teste e treino, em que:

$$\begin{aligned} &\textbf{Para, } i = 1 \textbf{ até } i = n - 1 \\ &\quad \textbf{Treino } (\textbf{ lote}_1, ..., \textbf{lote}_i \textbf{ }) \\ &\quad \textbf{Teste } (\textbf{ lote}_{i+1} \textbf{ }) \end{aligned} \tag{6.22}$$

Optou-se então por definir um total de 11 lotes, originando assim um conjunto de 10 sequências de teste e treino. Cada lote encerrará um total de 50 trilhos, sendo que aquando da fase de teste é inibida a capacidade de aprendizagem do classificador. No entanto, em paralelo com a fase de teste, é realizado o treino de um novo classificador, que será utilizado no teste seguinte.

De modo a avaliar o sistema em relação à classificação e previsão de eventos anormais, seleccionou-se uma região da cena observada, assinalada a vermelho na Figura 6.17. Esta região, que constitui um acesso a uma das salas do edifício, pretende simular um espaço restrito. Assim, um evento anormal ocorre sempre que um objecto atravesse esta área.

Figura 6.17. Selecção de zona restrita a pessoas (região representada a vermelho).

Durante as fases de teste foi possível observar em acção real o sistema de detecção e previsão de comportamentos. Para cada período de teste, sempre que a trajectória do objecto era reconhecida pelo classificador *DOG*, e caso não fosse previsível a violação da área previamente definida como restrita, então, a caixa delimitadora do objecto apresentaria a cor verde. Por outro lado, nos casos em que a trajectória era reconhecida como tendo uma elevada probabilidade de originar a violação do espaço restrito, então seria assinalada com a cor vermelha. Sempre que o classificador desconhecesse por completo a trajectória, a caixa delimitadora exibiria a cor amarela, sinalizando deste modo a ocorrência de um **evento não-usual**.

A Figura 6.18 apresenta uma sequência de imagens, retirada num período inicial de treino, onde após alguns segundos de permanência do objecto em cena, o classificador identifica um **evento não-usual**.

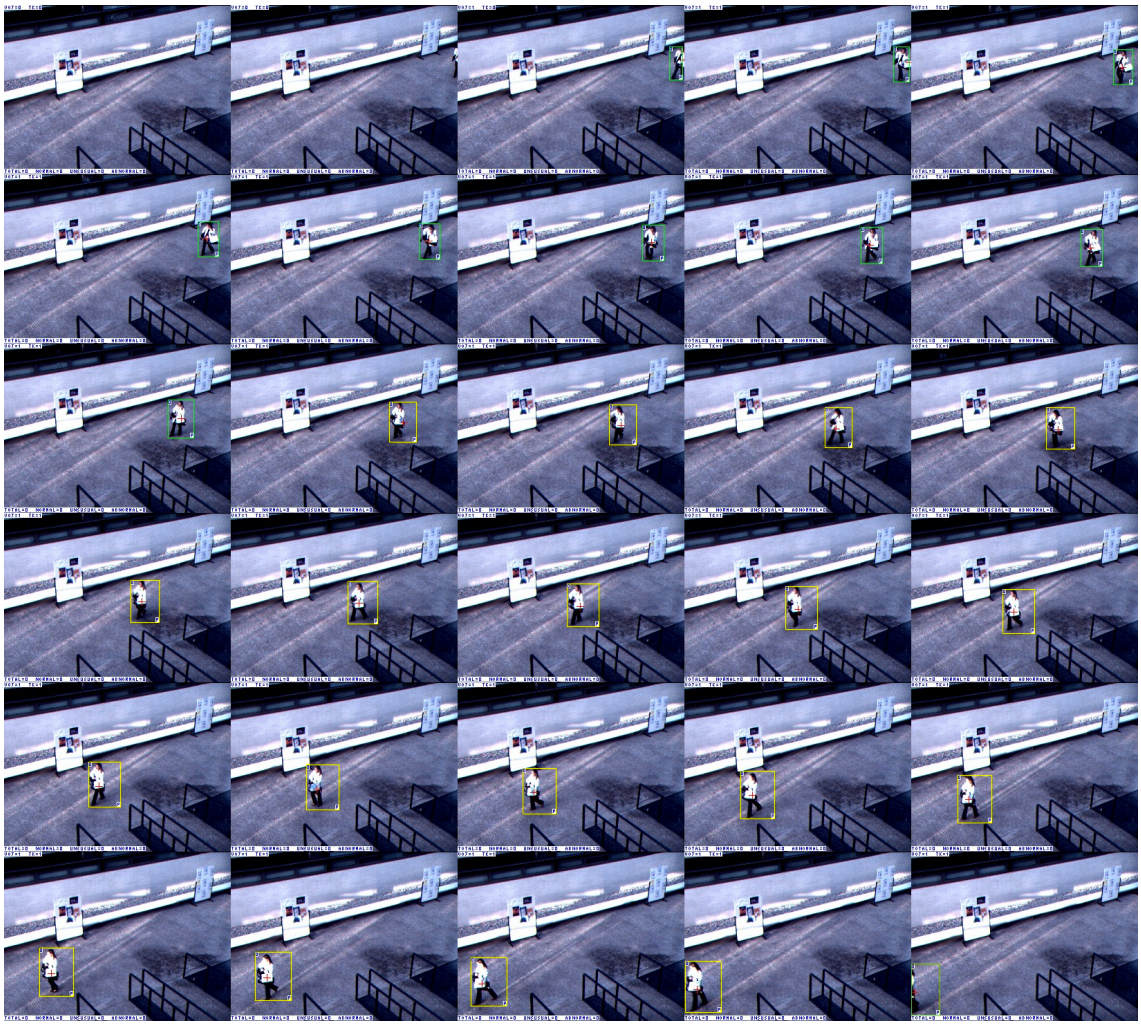


Figura 6.18. Sequência de imagens capturadas durante a passagem de um objecto.

A Figura 6.19 apresenta uma sequência de imagens, retiradas durante uma das fases de teste. Como se pode observar, o classificador reconhece ambos os trajectos, não considerando haver qualquer situação que leve à ocorrência de um **evento anormal**. Note-se ainda que o sistema realiza de forma eficaz o seguimento dos vários objectos presentes na cena.

Figura 6.19. Sequência de imagens.

Após a realização dos testes é então possível obter métricas de desempenho do classificador, como a curva *ROC* e valor da *AUC*. Outras informações como os mapas de dispersão dos objectos em cena, os trilhos mais frequentes, e as áreas de entrada e saída de cena são também identificadas.

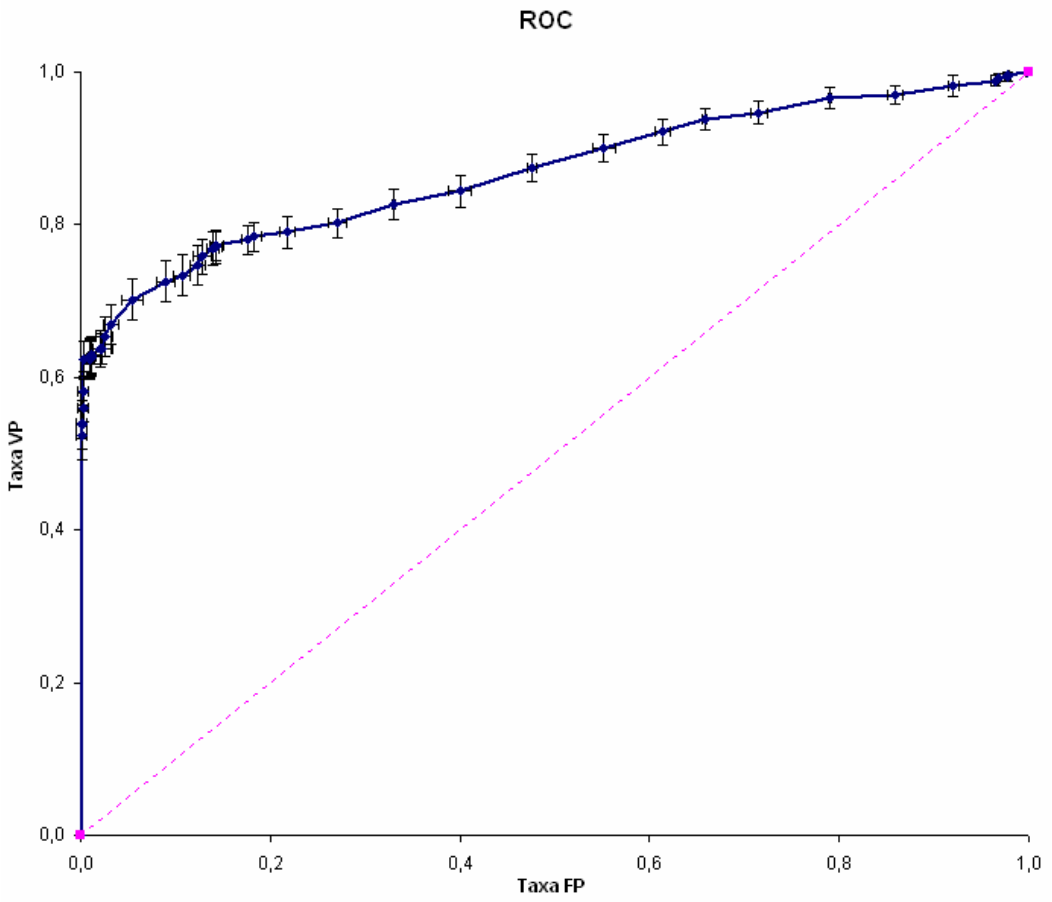


Figura 6.20. Curva *ROC* para o classificador, com apresentação das variações para um intervalo de confiança de 95%.

A curva *ROC* (Figura 6.20), obtida após o teste real, permite observar a variação da taxa de verdadeiros positivos (TP) de acordo com a taxa de falsos positivos (FP). Como se pode constatar, à medida que a taxa TP aumenta, aumenta também a taxa FP, embora numa relação não-linear. Para baixos valores de verdadeiros positivos, os resultados do classificador são pouco afectados por erros. Contudo, para taxas de TP mais elevadas verifica-se uma ocorrência mais frequente de falhas na classificação.

Analisando a *AUC* dos testes efectuados, observou-se que esta exibe um valor na ordem dos 87%. Confirma-se assim a avaliação levada a cabo com dados sintéticos onde o valor se situava em 86%.

Em relação às Taxas de Antecipação, observadas durante as sequências de treino e teste, é possível perceber o efeito que o aumento do número de dados disponibilizados para a aprendizagem do classificador causa na previsão da violação de um espaço restrito. Apesar de não se estabelecer uma relação directa e linear, constata-se que o classificador obtém uma melhoria das capacidades de detecção antecipada de eventos anormais com o aumento do número de trilhos utilizados no treino. Como seria de esperar, verifica-se também maior capacidade de previsão para um *Limite de Alarme* superior.

Tabela 6.2. Tabela com identificação, para cada teste, do número de trilhos que originaram uma violação do espaço restrito. Valores para 75%, 50% e 25% de *Limite de Alarme*.

Teste	Trilhos/Lote	Violação	75%	50%	25%
1	50	7	2	3	3
2	50	5	2	4	4
3	50	6	1	3	4
4	50	11	4	7	9
5	50	23	9	15	20
6	50	8	3	5	7
7	50	3	2	2	3
8	50	18	7	14	17
9	50	7	3	6	6
10	50	9	4	7	9

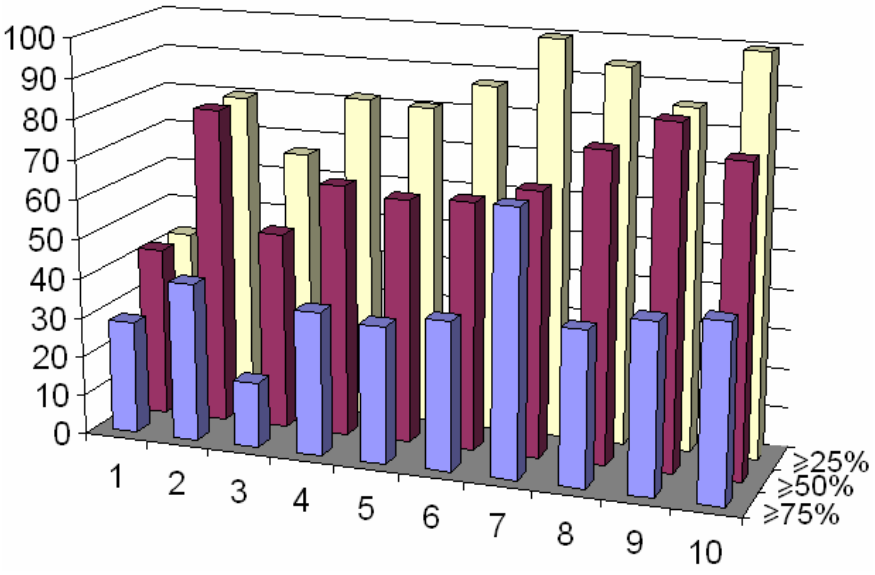


Figura 6.21. Gráfico de percentagem de acerto na previsão de eventos anormais. O eixo dos "x" representa cada teste realizado, o eixo dos "y" apresenta a percentagem de acerto, e o eixo dos "z" define os *Limites de Alarme* testados.

Note-se que durante a realização do teste foram observados 550 trilhos. Deste conjunto, 97 trilhos levaram, em algum segmento do seu trajecto, à violação da área considerada como restrita.

Finalmente, apresentam-se algumas informações relativas aos movimentos protagonizados pelos objectos, nomeadamente as áreas de entrada e saída de cena, bem como o trilho mais frequente. Este tipo de informação, apesar de não ter uma relação directa com a detecção de eventos anormais ou não-usuais, pode ser de grande utilidade para aqueles que têm a seu cargo a elaboração de planos e projectos de segurança para um determinado espaço.

Pela análise destes mapas contendo dados relativos à movimentação de pessoas, torna-se assim possível tomar decisões fundamentadas em dados reais. Tais dados constituem uma ferramenta importante para a identificação de regiões críticas e possivelmente para a reformulação dos procedimentos e das políticas de segurança de um dado local.

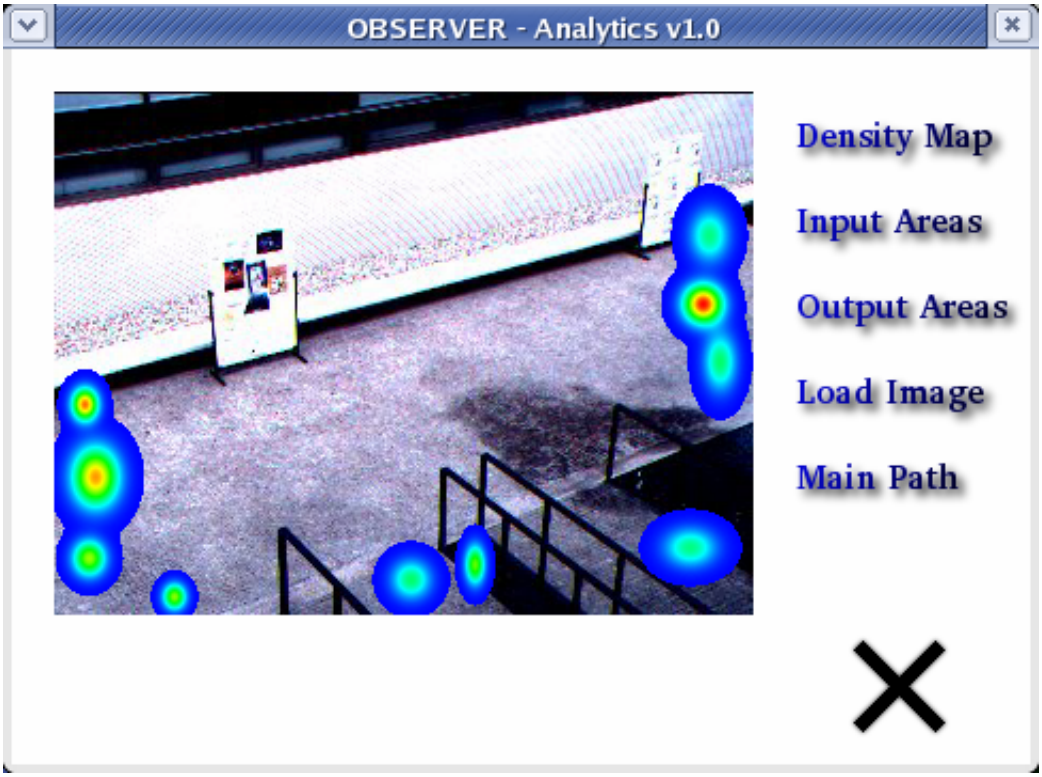


Figura 6.22. Identificação das regiões de entrada e saída de objectos da área monitorizada.

Como se pode constatar na Figura 6.22, o classificador identificou 10 regiões de entrada e saída de objectos em cena. A figura ilustra a relevância de cada área, pela exibição de um gradiente da cor vermelha para azul, onde o vermelho indica maior representatividade.

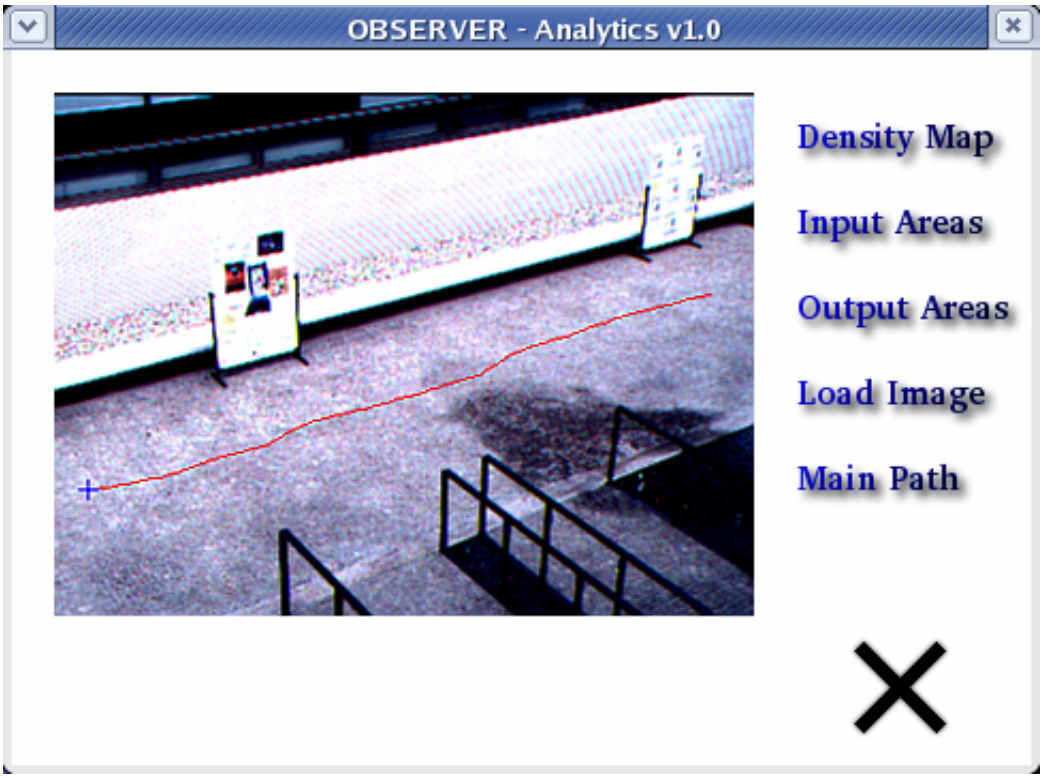


Figura 6.23. Representação do trilho principal, descrevendo o trajecto mais frequente.

Na Figura 6.23 apresenta-se o trilho mais frequentemente observado durante o teste do sistema. Esta informação pode ser útil em casos onde se pretendam estudar os fluxos de pessoas em edifícios ou ruas.

6.8. Discussão

No presente capítulo definiu-se que qualquer acção levada a cabo por um objecto (i.e. pessoa, grupo de pessoas ou veículo) pode ser classificada segundo um: evento **normal**, **não usual** ou **anormal**. Seguindo tal tipologia de classificação, entende-se como um **evento anormal** aquele que leva à violação de um espaço restrito. Por outro lado, definem-se como **eventos normais** aqueles que são frequentemente observados e que não originam, nem se prevê que venham a originar, uma violação de uma área restrita. Comportamentos até então desconhecidos pelo classificador são considerados **eventos não usuais**.

Tendo por base tal estruturação, propõe-se neste capítulo dois classificadores (*N-ary Trees* e *Dynamic Oriented Graph*) que possibilitam a detecção e previsão de eventos anormais, bem como facultam um meio de sinalização de comportamentos desconhecidos.

Tanto o classificador *N-ary Trees* como o *Dynamic Oriented Graph (DOG)* realizam uma aprendizagem não supervisionada à medida que os dados de treino vão sendo processados. Embora exibam uma estrutura bastante similar, o *DOG* distingue-se do seu antecessor, i.e. *N-ary Trees*, por realizar o ajuste dinâmico do modelo de classificação. O *DOG* proporciona uma adaptação contínua, i.e. exibe plasticidade, permitindo a geração de novas classes, assim como a sua reorganização caso os dados o determinem.

Através de uma análise comparativa entre os dois classificadores, utilizando para o efeito um conjunto de dados sintéticos, constatou-se a superioridade no desempenho do *DOG*. Este classificador apresenta superiores valores de *ROC* e *AUC*, contudo o seu maior benefício reside na elevada capacidade de processamento. Note-se que o classificador *DOG* requereu 1 minuto e 29 segundos para executar o teste de avaliação, em contraste com o classificador *N-ary Trees* cujo tempo de execução se situou à volta das 190 horas.

Tendo em consideração tais resultados, optou-se pela inclusão do classificador *DOG* na câmara inteligente. Por conseguinte, os testes reais tiveram como propósito a validação exclusiva deste classificador.

Em ambiente real, a avaliação do sistema foi realizada através de uma técnica de *re-treino incremental*, dado que se pretendia preservar a ordem pela qual os objectos se apresentavam em cena. Assim, definiu-se um esquema de 11 lotes, com 50 trilhos por lote, possibilitando

10 sequências de treino e teste. Seleccionou-se ainda como área restrita, uma região da área monitorizada.

Em testes preliminares, com o propósito de avaliar o desempenho do *hardware* da câmara de vídeo, constatou-se uma total incapacidade de ajuste automático da íris. Tal característica deve-se ao facto de se tratar de uma câmara desenvolvida para aplicações industriais e laboratoriais, onde esta é operada sob um ambiente controlado. Verificou-se ainda que o ajuste por *software* a sombras e brilhos não chega, por si só, para colmatar esta carência da câmara de vídeo.

Devido a tais condicionantes, o teste real foi executado de forma faseada, com a aquisição, treino e teste de um ou dois lotes por dia. Durante as fases de teste foi tido um especial cuidado em evitar períodos onde as condições meteorológicas originassem uma forte e súbita variação da intensidade luminosa. Lotes afectados por erros com origem na variação da intensidade luminosa não foram incluídos no teste. Nestes casos, o modelo de classificação anterior era reposto antes de se proceder a novo treino.

Os resultados obtidos com o teste real vieram confirmar a tese de possibilidade de antecipação automática da ocorrência de um evento anormal. A partir do modelo do classificador *DOG*, foi ainda possível retirar informação relativa a zonas de entrada e saída de pessoas, bem como do trilho mais frequente.

Capítulo 7

7. Conclusões

Esta tese encerra com uma breve recapitulação dos tópicos apresentados nas secções anteriores. Assim, num curto sumário, descrevem-se de forma sucinta os componentes que constituem o sistema desenvolvido para a detecção e previsão de comportamentos observados através de câmaras de vídeo. São enunciados os métodos propostos para a segmentação de objectos em movimento, detecção de sombras e brilhos, detecção e remoção de fantasmas, seguimento e classificação de objectos, bem como as técnicas utilizadas para a detecção e previsão de comportamentos.

Incluída ainda neste capítulo encontra-se uma listagem dos contributos alcançados no decorrer deste trabalho nas áreas de processamento de imagem, análise de imagem e inteligência artificial.

O capítulo cessa com a apresentação de possíveis direcções de investigação para trabalho futuro. Aí, são efectuadas algumas sugestões que visam complementar o trabalho realizado, bem como abrir novos caminhos de investigação.

7.1. Sumário da Tese

Este trabalho de doutoramento visa a investigação de técnicas e métodos que permitam detectar e prever comportamentos anómalos, passíveis de originar eventos de quebra de segurança em locais monitorizados por uma câmara de vídeo a cores, monocular e fixa. Pretendia-se realizar tal tarefa de aferição do tipo de comportamento, minimizando a necessidade de informação de contexto, i.e. sem modelação tridimensional da cena observada ou qualquer definição prévia de comportamentos conhecidos ou imagináveis.

Procedeu-se então ao fraccionamento do trabalho em três fases distintas: segmentação de objectos em movimento; seguimento de objectos ao longo do seu trajecto; e detecção e previsão de comportamentos. Tal método de organização, não só faculta uma possibilidade de validação intermédia das técnicas apresentadas, mas possibilita também futuros melhoramentos em qualquer destas áreas, reutilizando o trabalho desenvolvido nas restantes.

A fase inicial do trabalho foi dedicada à segmentação de movimento. Aí, estudaram-se técnicas que possibilitassem uma correcta detecção de objectos em deslocação pela cena. Para tal, realizou-se uma análise exaustiva de modelos de representação de cor, com o propósito de seleccionar aquele, ou aqueles, que melhor se adequariam aos requisitos de invariabilidade à intensidade da iluminação. Note-se que uma segmentação exacta e precisa do objecto, não afectada por erros com origem na sua sombra ou alterações da intensidade luminosa, irá facilitar e aumentar a precisão dos restantes processos situados a jusante.

Ainda nesta primeira fase, desenvolveu-se um método para a segmentação de movimento, composto por técnicas de detecção de fantasmas, detecção de contornos e adaptação do plano de fundo. A técnica proposta para a detecção e remoção de fantasmas constitui uma mais-valia, vindo colmatar uma fragilidade evidenciada pelas técnicas de segmentação baseadas em subtracção de plano de fundo.

Uma especial atenção deve ser dada à implementação da funcionalidade de adaptação do plano de fundo. Através da experimentação, constatou-se que deve ser mantida em memória uma imagem de plano de fundo em formato de vírgula flutuante. Apesar de este tipo de dados requerer maior espaço reservado em memória e a sua operação ser mais complexa, é o único tipo de dados que assegura o correcto ajuste da imagem de fundo. Aquando da utilização de um tipo de dados de 8 *bits*, como o tipo *unsigned char*, verificou-se

a existência de uma incapacidade de adaptação sempre que o valor do ponto da imagem actual se encontra sensivelmente próximo do valor do ponto da imagem de fundo.

Na segunda fase do trabalho abordou-se o tema do seguimento de objectos em movimento. Recorreu-se então à representação de um qualquer objecto por um *Modelo de Aparência*, constituído por uma *Máscara de Probabilidade* e *Imagem de Aparência*. O modelo, mantido em memória durante o período de tempo em que o objecto é observado em cena, permite identificar as características de forma e cor que discriminam o objecto alvo. O *Modelo de Aparência* é construído e mantido de forma dinâmica, assegurando adaptabilidade às alterações morfológicas protagonizadas pelo objecto. Esta capacidade de adaptação verifica-se ser particularmente útil e eficaz no seguimento de objectos deformáveis.

A classificação dos objectos em: pessoa, grupo, ou veículo, foi realizada com recurso às formas prováveis dos objectos, mantidas pelas *Máscaras de Probabilidade*. A utilização desta informação, em detrimento da forma definida pela região segmentada (do objecto) a cada instante, teve por base a estabilidade da primeira. Tal propriedade advém do facto das alterações morfológicas instantâneas serem assimiladas lentamente pela *Máscara de Probabilidade*. Assim, mesmo que o objecto altere momentaneamente a sua forma, esta mantém-se sensivelmente inalterada pelo modelo durante um determinado período de tempo. A corrente abordagem assegura ainda a correcta classificação nos casos de oclusão parcial do objecto.

A última fase deste trabalho de doutoramento foi dedicada à elaboração de técnicas que assegurassem a detecção e previsão de comportamentos pela identificação de padrões nos dados provenientes dos processos de segmentação e seguimento de objectos. Com efeito, foram propostos dois novos classificadores, nomeadamente o *N-ary Trees* e o *Dynamic Oriented Graph*.

O *N-ary Trees*, é um classificador não supervisionado, que necessita de uma fase de treino para adquirir um modelo dos padrões observados. Embora o classificador apresente resultados satisfatórios em termos de classificação de comportamentos, este manifesta um aspecto que se afigura o problemático na sua aplicação em ambiente real de vigilância. O facto do modelo gerado permanecer estático após a fase de treino, ou seja, não possuir capacidade de adaptação, impede que sejam incorporadas no classificador alterações de padrões que se venham posteriormente a observar no ambiente sob monitorização.

O *Dynamic Oriented Graph*, é um classificador que implementa o mesmo conceito que o classificador anterior, mas a aprendizagem não supervisionada é realizada de forma a possibilitar o ajuste dinâmico do modelo, de acordo com as alterações de padrões observadas ao longo do tempo. O *DOG* foi desenvolvido com a finalidade de proporcionar um modelo de classificação adaptativo, i.e. exibir plasticidade, permitindo a geração de novas classes, bem como possibilitar a sua reorganização caso os dados assim o determinem.

Através de testes com dados sintéticos, constatou-se a superioridade do classificador *DOG* em relação ao *N-ary Trees*. Por conseguinte, foi seleccionado o primeiro para incorporação no protótipo da câmara inteligente. Os testes reais, realizados no campus da Universidade do Minho, validaram a tese de possibilidade de identificação e previsão de **comportamentos anormais**, observados por uma câmara de vídeo.

7.2. Contributos

A detecção e previsão de comportamentos anormais, com vista ao auxílio da tarefa de vigia e manutenção da segurança de espaços públicos, tem vindo a despertar o interesse de um cada vez maior número de investigadores. Não obstante os importantes contributos dos trabalhos já realizados, muitos deles referenciados ao longo desta tese, algumas questões mereciam um aperfeiçoamento e, em outros casos, era relevante o estudo de técnicas alternativas nas áreas de processamento e análise de imagem, bem como no campo da inteligência artificial. Nesta tese apresentam-se contributos para as três áreas abrangidas.

De uma forma objectiva, esta tese contribui com:

(i) uma nova técnica de segmentação de objectos em movimento, baseada na combinação de subtracção de plano de fundo adaptativo com a diferença entre imagens (a técnica proposta permite segmentar objectos, removendo do resultado da segmentação erros relativos a sombras, brilhos e fantasmas) [Duque et al., 2005; 2008];

(ii) um conjunto de técnicas de extracção de dados, nomeadamente a classificação dos objectos observados em pessoas, grupos de pessoas e veículos [Duque et al., 2006b];

(iii) um algoritmo de seguimento de objectos deformáveis, que garante o seguimento durante a ocorrência de interacções complexas entre objectos, como são o caso da fusão, separação e oclusão de objectos [Duque et al., 2006b; 2008]; e

(iv) duas propostas originais para a identificação e previsão de padrões de actividade, baseadas em dois classificadores denominados por *N-ary Trees* e *Dynamic Oriented Graph* [Duque et al., 2006; 2007; 2007b].

7.3. Trabalho Futuro

No decorrer deste trabalho foram tomadas opções que determinaram o rumo da investigação. Apesar de se ter procurado fundamentar o mais possível cada escolha de um método ou técnica em detrimento de outras, diferentes opções poderiam ter sido tomadas, o que provavelmente originariam resultados distintos e levantariam outras questões de investigação.

Não se espera com esta tese apontar um caminho único para a detecção e previsão de comportamentos anormais, observados através de câmaras de vídeo, sendo que existem muitas outras vias que poderão conduzir a óptimos resultados. Na presente secção, são apresentadas algumas sugestões para o trabalho a desenvolver, que visam complementar o trabalho realizado, bem como abrir novos caminhos de investigação.

Assim, relativamente à detecção de objectos em movimento, sugere-se o desenvolvimento de técnicas que possibilitem a diminuição de falsos positivos originados por oscilações da câmara de vídeo. Apesar de em ambiente laboratorial tais oscilações praticamente não se verificarem, em situações reais, onde por vezes as câmaras se encontram colocadas no topo de postes de elevada estatura, este aspecto pode ser crucial para a viabilidade do sistema.

No que diz respeito à classificação de objectos, seria de interesse a implementação de diferentes métodos de classificação, nomeadamente através de técnicas de redução de dimensão como a *PCA* e a *MFA* (*Marginal Fisher Analysis*). Este estudo deveria ser acompanhado de uma análise do peso computacional de cada abordagem, tendo em vista a sua utilização num sistema tempo-real.

Sugere-se também o desenvolvimento de uma base de dados para recepção de meta-informação via *MPEG-7*, com indexação de eventos de vídeo-vigilância. Deveria ainda ser estudado um método para pesquisa e descoberta de eventos que, através da implementação

de algoritmos de *data mining* sobre os conteúdos da base de dados, permitisse a sinalização de comportamentos anormais ou padrões de comportamentos observados num período temporal mais alargado (e.g. dias, semanas ou meses).

Uma outra proposta de trabalho futuro, que se enquadra na filosofia de descentralização, consiste no desenvolvimento de uma plataforma de *grid computing* para o processamento e análise das imagens capturadas pelas câmaras de vídeo. A necessidade de experimentação de tal abordagem encontra fundamento no considerável número de computadores existentes nas instituições, e cujo processamento disponível poderia ser utilizado para processar imagens provenientes de câmaras de vídeo IP comuns.

Bibliografia

[Abrantes et al., 2002] A. Abrantes, J. Marques & J. Lemos, "Long term tracking using bayesian networks," *Proceedings of the International Conference on Image Processing*, vol. III, pp. 609-612, 2002.

[Ainsworth, 2004] T. Ainsworth, "CCTV – Performance testing," *Security OZ Magazine*, Issue 9, pp. 50-59, 2004.

[Andleigh & Thakrar, 1996] Prabhat K. Andleigh & Kiran Thakrar, "Multimédia Systems Design," *Prentice Hall*, 1996.

[APB, 2008] Associação Portuguesa de Bancos, "Dados sobre a Banca em Portugal relativos ao exercício de 2007 – Boletim Informativo," *Associação Portuguesa de Bancos*, Ano 21, Nº 41, Julho de 2008.

[Arsic et al., 2005] D. Arsic, F. Wallhoff, B. Schuller & G. Rigoll, "Video based online behavior detection using probabilistic multi-stream fusion," *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, pp. 606-609, 2005.

[Ayers & Shah, 2001] D. Ayers & M. Shah, "Monitoring human behavior from video taken in an office environment," *Image and Vision Computing*, Elsevier, vol. 19, n. 12, pp. 833-846, 2001.

[Bashir & Porikli, 2006] F. Bashir & F. Porikli, "Performance evaluation of object detection and tracking systems," *Proceedings of the 9th IEEE International Workshop on PETS*, New York, pp. 7-14, 2006.

[Blueeyevideo, 2006] Blueeyevideo, 2006, http://www.blueeyevideo.com

[Boult et al., 1998] T. Boult, A. Erkin, P. Lewis, R. Michaels, C. Power, C. Qian & W. Yin, "Frame-rate multi-body tracking for surveillance," *Proceedings of the DARPA Image Understanding Workshop*, pp. 305-308, 1998.

[Boult et al., 1999] T. Boult, R. Michaels, X. Gao, P. Lewis, C. Power, W. Yin & A. Erkan, "Frame-rate omnidirectional surveillance and tracking of camouflaged and occluded

targets," *Proceedings of the 2nd IEEE International Workshop on Visual Surveillance*, pp. 48-55, 1999.

[Boult et al., 2001] T. Boult, R. Michaels, X. Gao & M. Eckman, "Into the woods: visual surveillance of noncooperative and camouflaged targets in complex outdoor settings," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1382-1402, 2001.

[Bramberger et al., 2006] M. Bramberger, A. Doblander, A. Maier, B. Rinner & H. Schwabach, "Distributed embedded smart cameras for surveillance applications," *Computer – IEEE Computer Society*, vol. 39, n. 2, pp. 68-75, 2006.

[Burges, 1998] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Proceedings of the Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[Buxton & Gong, 1995] H. Buxton & S. Gong, "Visual surveillance in a dynamic and uncertain world," *Artificial Intelligence – Special Volume on Computer Vision*, vol. 78, no. 1-2, pp. 431-459, 1995.

[Buxton & Howarth, 1996] H. Buxton & R. Howarth, "Watching behaviour: the role of context and learning," *Proceedings of the International Conference on Image Processing*, Lausanne, pp. 797-800, 1996.

[Buxton, 1997] H. Buxton, "Advanced visual surveillance using bayesian networks," *IEE Colloquium on Image Processing for Security Applications*, pp. 9/1-9/5, 1997.

[Buzan, 2004] D. Buzan, "Robust tracking of human motion," *Masters Project Final Report*, Tech Report Nº2004-016, Boston University, Boston, 2004.

[Buzan et al., 2004] D. Buzan, S. Sclaroff & G. Kollios, "Extraction and clustering of motion trajectories in video," *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, pp. 521-524, 2004.

[Canny, 1986] J. Canny, "A computational approach to edge detection," *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 8, no. 6, pp. 679-714, 1986.

[CAVIAR, 2006] Context Aware Vision Using Image-based Active Recognition, CAVIAR 2006, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

[Chang & Ansari, 2005] C. Chang & R. Ansari, "Kernel particle filter for visual tracking," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 242-245, 2005.

[Chen et al., 2005] J. Chen, T.N. Pappas, A. Mojsilović & B.E. Rogowitz, "Adaptive perceptual color-texture image segmentation, " *Transactions on Image Processing*, IEEE, vol. 14, no. 10, pp. 1524-1536, 2005.

[Citilog, 2006] Citilog, 2006, http://www.citilog.fr/index_en.php

[Cohen & Li, 2003] I. Cohen & H. Li, "Inference of human postures by classification of 3D human body shape," *Proceedings of the IEEE International Workshop on Analysis and Modelling of Faces and Gestures*, pp. 74-81, 2003.

[Collins et al., 2000] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto & O. Hasegawa, "A system for video surveillance and monitoring," *VSAM Final Report*, Carnegie Mellon University, Pittsburgh, 2000.

[Cooley & Tukey, 1965] J.W. Cooley & J.W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation Journal*, vol. 19, n. 90, pp. 297-301, 1965.

[Cormen et al., 1990] T.H. Cormen, C.E. Leiserson & R.L. Rivest, *Introduction to Algorithms*, M.I.T. Press, 1990.

[Corrall, 1991] D. Corrall, "VIEWS: computer vision for surveillance applications," *IEE Colloquium on Active and Passive Techniques for 3-D Vision*, London, pp. 8/1-8/3, 1991.

[Cover & Hart, 1967] T.M. Cover & P.E. Hart, "Nearest Neighbor Pattern Classification," *Transactions on Information Theory*, IEEE, vol. 13, no. 1, pp. 21-27, 1967.

[Cucchiara et al., 2001] R. Cucchiara, C. Grana, G. Neri, M. Piccardi & A. Prati, "The Sakbot system for moving object detection and tracking," *Video-Based Surveillance Systems – Computer Vision and Distributed Processing*, Kluwer Academia, pp. 145-157, 2001.

[Cucchiara et al., 2001b] R. Cucchiara, C. Grana, M. Piccardi, A. Prati & S. Sirotti, "Improving shadow suppression in moving object detection with HSV color information," *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 334-339, 2001.

[Cucchiara et al., 2002] R. Cucchiara, C. Grana, G. Neri, M. Piccardi & A. Prati, "The Sakbot system for moving object detection and tracking," *Proceedings of the 10th ACM International Conference on Multimedia*, ACM Press, France, pp. 223-226, 2002.

[Cucchiara et al., 2004] R. Cucchiara, C. Grana & G. Tardini, "Track-based and object-based occlusion for people tracking refinement in indoor surveillance," *Proceedings of the ACM 2nd International Workshop on Video Surveillance & Sensor Networks*, New York, USA, pp. 81-87, 2004.

[Davalo, 1988] E. Davalo & P. Naim, *DARPA Neural Network Study*, AFCEA International Press, 1988.

[Dee & Hogg, 2004] H. Dee & D. Hogg, "Detecting inexplicable behaviour," *Proceedings of the British Machine Vision Conference*, pp. 477-486, 2004.

[Dempster et al., 1977] A.P. Dempster, N.M. Laird & D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, vol. 39, pp. 1-38, 1977.

[Dennett, 1987] D.C. Dennett, *The Intentional Stance*, Bradford Books, Cambridge, 1987.

[Doermann & Mihalcik, 2000] D. Doermann & D. Mihalcik, "Tools and techniques for video performance evaluation," *Proceedings of the International Conference on Pattern Recognition*, Barcelona, pp. 4167-4170, 2000.

[Donald, 1999] C.H.M. Donald, "Assessing the human vigilance capacity of control room operators," *Proceedings of the International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centers*, Bath, UK, pp. 7-11, 1999.

[Dowling, 2002] J.E. Dowling, "The retina: an approachable part of the brain," *Belknap Press*, 2002.

[Duda et al., 2001] R.O. Duda, P.E. Hart & D.G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.

[Duque et al., 2005] D. Duque, H. Santos & P. Cortez, "Moving Object Detection Unaffected by Cast Shadows, Highlights and Ghosts," *Proceedings of the IEEE International Conference on Image Processing*, Genoa, Italy, pp. 413-416, 2005.

[Duque et al., 2006] D. Duque, H. Santos & P. Cortez, "N-ary Trees Classifier," *Proceedings of the 3rd International Conference on Informatics in Control, Automation and Robotics*, Setúbal, Portugal, 2006.

[Duque et al., 2006b] D. Duque, H. Santos & P. Cortez, "The OBSERVER: An Intelligent and Automated Video Surveillance System," *Image Analysis and Recognition*, Lecture Notes in Computer Science - Springer, pp. 889-909, 2006.

[Duque et al., 2007] D. Duque, H. Santos & P. Cortez, "Prediction of Abnormal Behaviors for Intelligent Video Surveillance Systems," *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, Honolulu - Hawaii, USA, pp. 362-367, 2007.

[Duque et al., 2007b] D. Duque, H. Santos & P. Cortez, "Previsão de Eventos Anormais em Vídeo-vigilância," *Simpósio Doutoral em Inteligência Artificial*, Guimarães, Portugal, pp. 165-174, 2007.

[Duque et al., 2008] D. Duque, H. Santos & P. Cortez, "Avaliação de Desempenho na Segmentação e Seguimento de Objectos no Sistema Observer," *Congresso Luso-Moçambicano de Engenharia*, Maputo, Moçambique, 2008.

[ETISEO, 2006] Video Understanding Evaluation, ETISEO 2006, http://www.silogic.fr/etiseo

[Fastcom, 2006] Fastcom Technology, SFA - Smoke and Fire Alert, 2006, http://www.fastcom.ch/pages/security_sfa.html

[Fawcett, 2003] T. Fawcett, "ROC graphs: notes and practical considerations for data mining researchers," Tech report HPL-2003-4, HP Laboratories, Palo Alto, USA, 2003.

[Flexer, 1996] A. Flexer, "Statistical evaluation of neural network experiments: minimum requirements and current practice," *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, Cybernetics and Systems 96, pp. 1005-1008, 1996.

[Freeman, 1970] H. Freeman, "Boundary encoding and processing, in Picture processing and psychopictorics," *Academic Press*, New York, pp. 241-266, 1970.

[Fukusima, 1997] S.S. Fukusima, "Sombras como indicadores da percepção de profundidade," *Psicologia Reflexão e Crítica*, vol. 10, n. 2, pp. 289-300, 1997.

[Funka-Lea & Bajcsy, 1995] G. Funka-Lea & R. Bajcsy, "Combining color and geometry for the active, visual recognition of shadows," *Proceedings of the 5th International Conference on Computer Vision*, pp. 203-209, 1995.

[Gevers & Smeulders, 1999] T. Gevers & A.W.M. Smeulders, "Color based object recognition," *Pattern Recognition*, vol. 32, n. 3, pp. 453-464, 1999.

[Ghidary et al., 2000] S.S. Ghidary, Y. Nakata, T. Takamori & M. Hattori, "Human detection and localization at indoor environment by home robot," *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, vol. 2, pp. 1360-1365, 2000.

[Giachetti et al., 1998] A. Giachetti, M. Campani & V Torre, "The use of optical flow for road navigation," *Transactions on Robotics and Automation*, IEEE, vol. 14, n. 1, pp. 34-48, 1998.

[Gill, 1962] A. Gill, "Introduction to the Theory of Finite-state Machines," McGraw-Hill, New York, 1962.

[Grimson et al., 1998] W.E.L. Grimson, C. Stauffer, R. Romano & L. Lee, "Using adaptive tracking to classify and monitor activities in a site," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 22-29, 1998.

[Hall et al., 2005] D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R. Fisher, J. Victor & J. Crowley, "Comparison of target detection algorithms using adaptive background models," *Proceedings of the 2nd Joint International Workshop on Visual Surveillance and Performance Evaluation of Tracking Surveillance*, Beijing, 2005.

[Haritaoglu et al., 1998] I. Haritaoglu, D. Harwood & L. Davis, "W4: Who? When? Where? A real time system for detecting and tracking people," *Proceedings of the International Conference on Face and Gesture Recognition*, pp. 222-227, 1998.

[Haritaoglu et al., 2000] I. Haritaoglu, D. Harwood & L. Davis, "W4: real-time surveillance of people and their activities," *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 22, n. 8, pp. 809-830, 2000.

[Herodotou et al., 1998] N. Herodotou, K.N. Plataniotis & A.N. Venetsanopoulos, "A color segmentation scheme for object-based video coding," *Proceedings of the IEEE Symposium on Advances in Digital Filtering and Signal Processing*, pp. 25-29, 1998.

[Hongeng et al., 2000] S. Hongeng, F. Bremond & R. Nevatia, "Representation and optimal recognition of human activities," *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1818-1825, 2000.

[Hongeng et al., 2004] S. Hongeng, R. Nevatia & F. Bremond, "Video-based event recognition: activity representation and probabilistic recognition methods," *Computer Vision and Image Understanding*, Elsevier, vol. 96, pp. 129-162, 2004.

[Horn & Schunck, 1981] B.K.P. Horn & B.G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185-203, 1981.

[Howarth & Buxton, 1992] R.J. Howarth & H. Buxton, "Analogical representation of spatial events for understanding traffic behaviour," *Proceedings of the 10th European Conference on Artificial Intelligence*, Vienna, pp. 785-789, 1992.

[Howarth & Buxton, 1998] R. Howarth & H. Buxton, "Attentional control for visual surveillance," *Proceedings of the IEEE Workshop on Visual Surveillance*, pp. 86-93, 1998.

[Hyder et al., 2002] Anthony K. Hyder, E. Shahbazian & Edward Waltz, "Multisensor Fusion," *Proceedings of the NATO Advanced Study Institute on Multisensor Data Fusion*, pp. 543, 2002.

[i-LIDS, 2006] Home Office, Science, Research & Statistics, i-LIDS 2006, http://scienceandresearch.homeoffice.gov.uk/hosdb

[Isard & Blake, 1996] M. Isard & A. Blake, "Contour tracking by stochastic propagation of conditional density," *Proceedings of the European Conference on Computer Vision*, pp. 343-356, 1996.

[Jack, 2005] K. Jack, "Video demystified – a handbook for the digital engineer," *Elsevier*, 4th Edition, 2005.

[Jaynes et al., 2002] C. Jaynes, S. Webb, R. Steele & Q. Xiong, "An open development environment for evaluation of video surveillance systems," *Proceedings of the 3rd International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen, pp. 32-39, 2002.

[Johnson & Hogg, 1996] N. Johnson & D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image and Vision Computing*, vol. 14, n. 8, pp. 609-615, 1996.

[Johnson & Hogg, 2002] N. Johnson & D. Hogg, "Representation and synthesis of behaviour using Gaussian mixtures," *Image and Vision Computing*, vol. 20, n. 12, pp. 889-894, 2002.

[Jolliffe, 1986] I.T. Jolliffe, "Principal Component Analysis," *Springer Series in Statistics*, Springer-Verlag, New-York, 1986.

[Jorge et al., 2004] P.M. Jorge, A.J. Abrantes & J. Marques, "On-line object tracking with bayesian networks," *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Systems*, Lisboa, 2004.

[Keogh & Ratanamahatana, 2005] E. Keogh & C.A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, Springer-Verlag, vol. 7, n. 3, pp. 358-386, 2005.

[Khalid & Naftel, 2005] S. Khalid & A. Naftel, "Classifying spatiotemporal object trajectories using unsupervised learning of basis function coefficients," *Proceedings of the 3rd ACM International Workshop on Video Surveillance & Sensor Networks*, pp. 45-52, 2005.

[Kim & Hwang, 2001] C. Kim & J.N. Hwang, "Video object extraction for object-oriented applications," *Journal of VLSI Signal Processing*, Kluwer Academic Publishers, vol. 29, pp. 7-21, 2001.

[Knill et al., 1997] D.C. Knill, P. Mamassian & D. Kersten, "Geometry of shadows," *Journal of the Optical Society of America*, vol. 14, n. 12, pp. 3216-3232, 1997.

[Kohavi, 1995] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, San Francisco, pp. 1137-1143, 1995.

[Kohonen, 1997] T. Kohonen, *Self-Organizing Maps*, 2nd ed., vol. 30, Springer-Verlag, New York, 1997.

[Kries, 1902] V. Kries, "Chromatic Adaptation," *Festschrift der Albrecht-Ludwigs-Universitt*, pp. 145-148, 1902. (Tradução: D.L. MacAdam, "Colorimetry-Fundamentals," *SPIE Milestone Series*, vol. MS 77, 1993.)

[Kruger et al., 1995] W. Kruger, W. Enkelmann & S. Rossle, "Real-time estimation and tracking of optical flow vectors for obstacle detection," IEEE Intelligent vehicle symposium, pp. 304-309, 1995.

[Lafferty et al., 2001] J. Lafferty, A. McCallum & F. Pereira, "Conditional random fields: probabilistic models for segmenting and labelling sequence data," *Proceedings of the 8th International Conference on Machine Learning*, pp. 282-289, 2001.

[Leo et al., 2004] M. Leo, T. D'Orazio, P. Spagnolo & A. Distante, "Complex human activity recognition for monitoring wide outdoor environments," *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 4, pp. 913-916, 2004.

[Leo et al., 2004b] M. Leo, T. D'Orazio & P. Spagnolo, "Human activity recognition for automatic visual surveillance of wide areas," *Proceedings of the 2nd International Workshop on Video Surveillance & Sensor Networks*, pp. 124-130, 2004.

[León & Sucar, 2002] R.D. León & L.E. Sucar, "Continuous activity recognition with missing data," *Proceedings of the 16th International Conference on Pattern Recognition*, vol. 1, pp. 439-442, 2002.

[Linde et al., 1980] Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design," *Transactions on Communications*, IEEE, vol. 28, n. 1, pp. 84-95, 1980.

[Lipton et al., 1998] A.J. Lipton, H. Fujiyoshi & R.S. Patil, "Moving target classification and tracking from real-time video," *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision*, pp. 8-14, 1998.

[Lucchese & Mitra, 1998] L. Lucchese & S.K. Mitra, "Na algorithm for unsupervised color image segmentation," *2nd Workshop on Multimedia Signal Processing*, IEEE, pp. 33-38, 1998.

[Lukac & Plataniotis, 2007] Rastislav Lukac & Konstantinos N. Plataniotis, "Color image processing: methods and applications," *CRC Press*, pp. 205, 2007.

[MacQueen, 1967] J. MacQueen, "Some methods for classifications and analysis of multivariate observations," *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, vol. 1, pp. 281-297, 1967.

[Makris & Ellis, 2001] D. Markis & T. Ellis, "Finding paths in video sequences," *Proceedings of the British Machine Vision Conference*, vol. 1, pp. 263-272, 2001.

[Makris & Ellis, 2002] D. Markis & T. Ellis, "Path detection in video surveillance," *Image and Vision Computing Journal*, Elsevier, vol. 20, pp. 895-903, 2002.

[Makris & Ellis, 2003] D. Makris & T. Ellis, "Automatic learning of an activity-based semantic scene model," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 183-188, 2003.

[Makris & Ellis, 2005] D. Makris & T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *Transactions on Systems, Man, And Cybernetics – Part B: Cybernetics*, IEEE, vol. 35, n. 3, pp. 397-408, 2005.

[Mamassian et al., 1998] P. Mamassian, D.C. Knill & D. Kersten, "The perception of cast shadows," *Trends in Cognitive Sciences*, Elsevier, vol. 2, n. 8, pp. 288-295, 1998.

[Matsushita et al., 2004] Y. Matsushita, K. Nishino, K. Ikeuchi & M. Sakauchi, "Illumination normalization with time-dependent intrinsic images for video surveillance," *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 26, n. 10, pp. 1336-1347, 2004.

[McCallum et al., 2000] A. McCallum, D. Freitag & F. Pareira, "Maximum entropy markov models for information extraction and segmentation," *Proceedings of the 17th International Conference on Machine Learning*, pp. 591-598, 2000.

[McLachlan & Basford, 1988] G.J. McLachlan & K.E. Basford, "Mixture models: inference and applications to clustering," *Statistics, textbooks and monographs*, vol. 84, 1988.

[Mecocci et al., 2003] A. Mecocci, M. Pannozzo & A. Fumarola, "Automatic detection of anomalous behavioural events for advanced real-time video surveillance," *Proceedings of the International Symposium on Computational Intelligence for Measurement Systems and Applications*, pp. 187-192, 2003.

[Medioni et al., 2001] G. Medioni, I. Cohen, F. Brémond, S. Hongeng & R. Nevatia, "Event detection and analysis from video streams," *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 23, n. 8, pp. 873-889, 2001.

[Metsis et al., 2006] V. Metsis, I. Androutsopoulos & G. Paliouras, "Spam filtering with naïve bayes – Which naïve bayes?," *Proceedings of the 3rd Conference on Email and Anti-Spam*, California, USA, 2006.

[Milenova & Campos, 2002] B.L. Milenova & M.M. Campos, "O-cluster: scalable clustering of large high dimensional data sets," *Proceedings of the 2nd IEEE International Conference on Data Mining*, pp. 290-297, 2002.

[Moreno et al., 2001] F. Moreno, J. Andrade-Cetto & A. Sanfeliu, "Localization of human faces fusing color segmentation and depth from stereo," *Proceesings of the 8th IEEE International Conference on Emerging Technologies and Factory Automation*, vol. 2, pp. 527-535, 2001.

[Muneeswaran et al., 2006] K. Muneeswaran, L. Ganesan, S. Arumugam & K.R. Soundar, "Texture image segmentation using combined features from spatial and spectral distribution, " *Pattern Recognition Letters*, Elsevier Science, vol. 27, no. 7, pp. 755-764, 2006.

[Nascimento et al., 2005] J.C. Nascimento, M.A.T. Figueiredo & J.S. Marques, "Segmentation and classification of human activities," *HAREM 2005 – International Workshop on Human Activity Recognition and Modelling*, Oxford, UK, 2005.

[Nascimento & Marques, 2004] J. Nascimento & J.S. Marques, "New performance evaluation metrics for object detection algorithms," *PETS 2004 – IEEE Workshop on Performance Evaluation of Tracking and Surveillance,* 2004.

[Naylor, 2006] M. Naylor, ADVISOR project website (2006), URL: http://www-sop.inria.fr/orion/ADVISOR

[NTSC, 1954] National Television Systems Committee, "NTSC signal specifications," *Proceedings of the IRE*, vol. 42, pp. 17-19, 1954.

[Otsu, 1979] N. Otsu, "A threshold selection method from gray-level histograms," *Transactions on Systems, Man and Cybernetics*, IEEE, vol. SMC-9, pp- 62-66, 1979.

[PETS, 2001] Performance Evaluation of Tracking and Surveillance Workshop, 2001, http://pets2001.visualsurveillance.org

[PETS, 2001b] Performance Evaluation of Tracking and Surveillance Workshop, 2001, http://www.cvg.cs.rdg.ac.uk/PETS2001/ANNOTATION/

[PETS, 2005] Performance Evaluation of Tracking and Surveillance Workshop, 2005, http://pets2005.visualsurveillance.org/

[PETS, 2006] Performance Evaluation of Tracking and Surveillance On-line Evaluation Service, 2006, http://www.cvg.cs.rdg.ac.uk/cgi-bin/PETSMETRICS/page.cgi?dataset

[Polidorio et al., 2003] A.M. Polidorio, F.C. Flores, N.N. Imai, A.M.G. Tommaselli & C. Franco, "Automatic shadow segmentation in aerial color images," *Proceedings of the 16th Brazilian Symposium on Computer Graphics and Image Processing*, pp. 270-277, 2003.

[Power & Schoonees, 2002] P.W. Power & J.A. Schoonees "Understanding background mixture models for foreground segmentation," *Proceedings of Image and Vision Computing*, pp. 267-271, 2002.

[Prewitt, 1970] J.M.S. Prewitt, "Object enhancement and extraction," *Picture Analysis and Psychopictorics*, Academic Press, 1970.

[Pun, 1980] T. Pun, "A new method for gray-level Picture thresholding using the entropy of the histogram," *Signal Processing*, vol. 2, pp. 223-237, 1980.

[Rabiner et al., 1978] L.R. Rabiner, A.E. Rosenberg & S.E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *Transactions on Acoustics, Speech, And Signal Processing*, IEEE, vol. 26, n. 6, pp. 575-582, 1978.

[Rabiner, 1989] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, n. 2, pp. 257-286, 1989.

[Regazzoni et al., 2001] C. Regazzoni, V. Ramesh & G.L. Foresti, "Scanning the issue/technology," *Proceedings of the IEEE*, vol. 89, n. 10, pp. 1355-1367, 2001.

[Remagnino & Jones, 2001] P. Remagnino & G.A. Jones, "Classifying surveillance events from attributes and behaviour," *Proceedings of the British Machine Vision Conference*, Manchester, pp. 685-694, 2001.

[Rumelhart & Zipser, 1986] D.E. Rumelhart & D. Zipser, "Feature discovery by competitive learning," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, vol.1, pp.151-193, 1986.

[Sacchi et al., 1999] C. Sacchi, C.S. Regazzoni & C. Dambra, "Remote cable-based video surveillance applications: the AVS-RIO project," *Proceedings of the 10th International Conference on Image Analysis and Processing*, Venice, Italy, pp. 1214-1215, 1999.

[Sakoe & Chiba, 1978] H. Sakoe & S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Transactions on Acoustics, Speech, and Signal Processing*, IEEE, vol. 26, n. 1, pp. 43-49, 1978.

[Salvador et al., 2001] E. Salvador, A. Cavallaro & T. Ebrahimi, "Shadow identification and classification using invariant color models," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1545-1548, 2001.

[Salvador & Ebrahimi, 2002] E. Salvador & T. Ebrahimi, "Cast shadow recognition in color images," *Proceedings of the 11th European Signal Processing Conference*, vol. 3, pp. 555-558, 2002.

[Salvador et al., 2003] E. Salvador, A. Cavallaro & T. Ebrahimi, "Spatio-temporal shadow segmentation and tracking," *Proceedings of the International Society for Optical Engineering*, vol. 5022, pp. 389-400, 2003.

[Salvador, 2004] E. Salvador, "Shadow segmentation and tracking in real-world conditions," *Thèse nº 3076*, École Polytechnique Fédérale de Lausanne, Lausanne, 2004.

[Sarabandi et al., 2004] P. Sarabandi, F. Yamazaki, M. Matsuoka & A. Kiremidjian, "Shadow detection and radiometric restoration in satellite high resolution images," *Proceedings of the International Geoscience and Remote Sensing Symposium*, vol. 6, pp. 3744-3747, 2004.

[Schlogl et al., 2004] T. Schlogl, C. Beleznai, M. Winter & H. Bischof, "Performance evaluation metrics for motion detection and tracking," *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 4, pp. 519-522, 2004.

[Senior et al., 2001] A.Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti & R Bolle, "Appearance models for occlusion handling," *Proceedings of the 2nd IEEE International Workshop on Performance Evaluation of Tracking in Surveillance*, Kauai, Hawaii, 2001.

[Senior, 2002] A. Senior, "Tracking people with probabilistic appearance models," *Proceedings of the International Workshop on Performance Evaluation of Tracking Surveillance Systems*, pp. 48-55, 2002.

[Shafer, 1985] S.A. Shafer, "Using color to separate reflection components," *Color Research and Applications*, vol. 10, pp. 210-218, 1985.

[Shannon, 1949] C.E. Shannon, "Communication in the presence of noise," *Proceedings of the Institute of Radio Engineers*, vol. 37, no. 1, pp. 10-21, 1949.

[Shin et al., 2001] M.C. Shin, D.B. Goldgof, K.W. Bowyer & S. Nikiforou, "Comparison of edge detection algorithms using a structure from motion task," *Transactions on Systems, Man, and Cybernetics*, IEEE, vol. 31, no. 4, pp. 589-601, 2001.

[Shioara et al., 1993] M. Shioara, H. Egawa, S. Sasaki, M. Nagle, P. Sobey & M.V. Srinivasan, "Real-time optical flow processor ISHTAR," *Proceedings of the Asian Conference on Computer Vision*, pp. 790-793, 1993.

[Sobel, 1978] I. Sobel, "Neighbourhood coding of binary images for fast contour following and general array binary processing," *Computer Graphics and Image Processing*, vol. 8, pp. 127-135, 1978.

[Stauffer & Grimson, 1999] C. Stauffer & W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246-252, 1999.

[Stauffer & Grimson, 2000] C. Stauffer & W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, vol. 22, n. 8, pp. 747-757, 2000.

[Survision, 2006] Survision, 2006, http://www.survision.fr/solutions/ac.php

[Suzuki et al., 2000] A. Suzuki, A. Shio, H. Arai & S. Ohtsuka, "Dynamic shadow compensation of aerial images based on color and spatial analysis," *Proceedings of the 15th International Conference on Pattern Recognition*, vol. 1, pp. 317-320, 2000.

[Theodoridis & Koutroumbas, 1999] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, San Diego, 1999.

[Tickner & Poulton, 1972] A.H. Tickner, E.C. Poulton, A.K. Copeman & D.C.V. Simmonds "Monitoring 16 television screens showing little movement," *Ergonomics*, Vol. 15, pp. 279-291, 1972.

[Tickner & Poulton, 1973] A.H. Tickner & E.C. Poulton, "Monitoring up to 16 synthetic television pictures showing a great deal of movement," *Ergonomics*, vol. 16, n. 4, pp. 381-401, 1973.

[Trias, 2005] A. Trias, "Shadow detection in videos acquired by stationary and moving cameras," *Master Thesis*, University of Maryland, 2005.

[Troscianko et al., 2004] T. Troscianko, A. Holmes, J. Stillman, M. Mirmehdi, D. Wright & A. Wilson, "What happens next? The predictability of natural behaviour viewed through CCTV cameras," *Perception*, Vol. 33, pp. 87-101, 2004.

[Vlachos et al., 2002] M. Vlachos, G. Kollios & D. Gunopulos, "Discovering similar multidimensional trajectories," *Proceedings of the 18th International Conference on Data Engineering*, pp. 673-684, 2002.

[Wallace, 1989] R. Wallace, "Finding natural clusters through entropy minimization," *PhD Thesis*, Carnegie Mellon University, CMU-CS-89-183, 1989.

[Wang & Lin, 2003] Y. Wang & Y. Lin, "A real-time approach for classification and tracking of multiple moving objects," *Proceedings of the 16th IPPR Conference on Computer Vision, Graphics and Image Processing*, pp. 67-74, 2003.

[Wesolkowski, 1999] S.B. Wesolkowski, "Color image edge detection and segmentation: a comparison of the vector angle and the Euclidean distance color similarity measures," *Master Thesis of Applied Science in Systems Design Engineering*, University of Waterloo, Canada, 1999.

[Williamson, 1996] J.R. Williamson, "Gaussian ARTMAP: a neural network for fast incremental learning of noisy multidimensional maps," *Neural Networks*, Elsevier Science, vol. 9, n. 5, pp. 881-897, 1996.

[Xu et al., 2003] M. Xu, L.Y. Duan, C.S. Xu & Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 189-192, 2003.

[Xu & Ellis, 2001] M. Xu & T. Ellis, "Illumination-invariant motion detection using colour mixture models," *Proceedings of the British Machine Vision Conference*, Manchester, UK, pp. 163-172, 2001.

[Xu & Zhang, 2001] Y. Xu & J. Zhang, "Abstracting human control strategy in projecting light source," *Transactions on Information Technology in Biomedicine*, IEEE, vol. 5, n. 1, pp. 27-31, 2001.

[Yacoob & Black, 1999] Y. Yacoob & M.J. Black, "Parameterized modelling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, n. 2, pp. 232-247, 1999.

[Yamamoto et al., 1995] S. Yamamoto, Y. Mae, Y. Shirai & J. Miura, "Realtime multiple object tracking based on optical flows," *Proceedings of the IEEE International Conference on Robotics and Automation*, Nagoya, Japan, pp. 2328-2333, 1995.

[Yoneyama et al., 2003] A. Yoneyama, C.H. Yeh & C.-C.J. Kuo, "Moving cast shadow elimination for robust vehicle extraction based on 2D joint vehicle/shadow models," *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 229-236, 2003.

[Young & Ferryman, 2005] D.P. Young & J.M. Ferryman, "PETS metrics: on-line performance evaluation service," *Proceedings of the 2nd Joint International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 317-324, 2005.

[Yue et al., 2003] Z. Yue, L. Zhao & R. Chellappa, "View synthesis of articulating humans using visual hull," *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 489-492, 2003.

[Zhang & Lu, 2001] D. Zhang & G. Lu, "Segmentation of moving objects in image sequence: a review," *Circuits, Systems, and Signal Processing*, vol. 20, no. 2, pp. 143-183, 2001.

[Zhong & Shi, 2003] H. Zhong & J. Shi, "Finding (un)usual events in video," *Technical Report CMU-RI-TR-03-05*, Carnegie Mellon University, 2003.

[Zhong et al., 2004] H. Zhong, J. Shi & M. Visontai, "Detecting unusual activity in video," *Proceeding of the Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 819-826, 2004.

[Zhu et al., 2006] G. Zhu, C. Xu, W. Gao & Q Huang, "Action recognition in broadcast tennis video using optical flow and support vector machine," *Proceedings of the European Conference on Computer Vision – Workshop on HCI*, pp. 89-98, 2006.