INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Análise de actividades humanas: interacção entre pessoas

João Pedro Saraiva Cabral Fernandes
(Licenciado)

Dissertação para a obtenção do Grau de Mestre em
Engenharia Electrotécnica e de Computadores

Júri:

Presidente:	Carlos Silvestre
Orientador:	Jorge Salvador Marques
Co-orientador:	Jacinto Nascimento
Vogal:	Fernando Pereira

Abril de 2011

Resumo

Esta tese apresenta um sistema capaz de identificar automaticamente interacções entre duas pessoas, recorrendo a uma câmera de video. É feita ainda uma comparação entre a utilização de dois modelos do corpo humano: modelo da silhueta e modelo anatómico.

O sistema implementado segue regiões activas numa sequência de imagens de cor para estimar o modelo da silhueta. De seguida este modelo é refinado para se obter o modelo anatómico. Esta operação é ralizada utilizando um modelo estatístico de cor e posição baseado em misturas Gaussianas para classificar pixeis individuais. De seguida os pixeis são agrupados em blobs que partilham características espaciais e de cor. Finalmente é utilizado um modelo grosseiro do corpo humano para classificar os blobs em partes anatómicas e se obter o modelo anatómico.

Após se estimar os modelos do corpo humano o sistema extraí características descritívas de cada um deles, que são utilizadas em conjunto com um classificados de K vizinhos mais próximos para se proceder à identificação das interacções.

O sistema foi testado para identificar quatro tipos de interacção: abraço, cruzamento, cumprimento e luta. As experiências mostram que o sistema implementado pode identificar interacções, apresentado uma elevada taxa de reconhecimento para ambos os modelos do corpo humano testados.

Palavras Chave: Sistemas de Video Vigilância, Interacções Humanas, Modelo da Silhueta, Modelo Anatómico.

Abstract

This thesis presents a system that is capable of automatically identify interactions between two persons using a video camera. Two models are used to represent the human body: the silhouette model and an anatomic model.

The implemented system tracks active regions in a color image sequence in order to estimate the silhouette model. This model is then refined to obtain the anatomic model. This is made at the expense of addicional processing steps, where Gaussian mixture models are used to classify individual pixels, based on their color. Next, the pixels are merged into blobs that share spacial and color features. Finally it is used a coarse human body model to classify the blobs into body parts, thus obtaining the anatomic model.

After estimating both of the human body models, the system extracts descriptive features from each one, that are used in conjunction with a k-nearest neighbors classifier to proceed to the interactions identification.

The system was tested to identify four types of interactions: hug, shake-hands, cross and fight. The experiments show that the implemented system can effectively identify interactions, presenting a high recognition rate for both of the human body models tested.

Keywords: Video Vigilance Systems, Human Interactions, Silhouette Model, Anatomic Model.

Índice

Capítulo 1

Introdução

1.1 Motivação

No mundo actual, os sistemas de vídeo vigilância estão omni presentes. Onde quer que vamos existem cameras de vigilância a observar-nos. Ao longo dos tempos, os sistemas de vídeo vigilância têm sido alvo de desenvolvimentos que permitem o desempenho de tarefas cada vez mais sofisticadas, e hoje em dia cada um de nós espera que estes sistemas actuem de forma eficaz na manutenção da nossa segurança.

As primeiras aplicações de vídeo vigilância datam dos anos 40 do século 20. O primeiro sistema CCTV, do inglês *Closed Circuit Television*, foi instalado pela Siemens AG em 1942 na Alemanha, e destinava-se a observar o lançamento dos mísseis V-2 a uma distância segura. Durante muitos anos, até ao aparecimento das vídeo cassetes, a utilização deste tipo de sistema restringiu-se à monitorização ao vivo, de imagens adquiridas em cameras estrategicamente colocadas.

Com o aparecimento da tecnologia de gravação analógica, a utilização dos sistemas CCTV ganhou popularidade, uma vez que possibilitou o armazenamento de dados para posterior análise. A tecnologia analógica apresentava no entanto alguns problemas pois as vídeo cassetes apresentavam uma capacidade de armazenamento de dados limitada, necessitando de ser trocadas todos os dias. Este problema foi minimizado nos anos 90 com o aparecimento dos multiplexadores digitais que permitiam, por exemplo, a gravação de dados apenas quando existia movimento, o que veio economizar bastante a quantidade de cassetes utilizadas. Adicionalmente permitiam ainda a gravação simultânea de várias cameras.

O avanço seguinte deu-se quando se trocou a gravação analógica pela gravação digital. Esta troca de formato permitiu que fosse possível a compressão de dados e o seu posterior armazenamento em discos rígidos, facto que acabou com os problemas de armazenamento. Adicionalmente, as imagens de vídeo digital têm maior qualidade e podem ser manipuladas para por exemplo aumentar a nitidez ou ampliar pormenores de uma imagem. O desenvolvimento das tecnologias de transmissão de dados digitais e o aparecimento da internet, o vídeo digital possibilitou aos sistemas de vídeo vigilância estarem presentes no mundo todo e a sua monitorização poder ser feita em qualquer parte do globo.

Com o acontecimento do 11 de Setembro de 2001 assistiu-se a uma evolução no paradigma da vídeo vigilância. O cidadão comum tomou consciência que os sistemas existentes, de índole passiva, eram insuficientes na manutenção da segurança pública, havendo uma necessidade de mudança para sistemas activos capazes de actuarem como agentes preventivos. Nesse sentido, e como resposta a esta tomada de consciência, foram desenvolvidos programas que permitiram tornar a vídeo vigilância mais eficaz como, re-

conhecedores faciais e métodos automáticos de seguimento de pessoas. Hoje em dia já não interessa só saber quantas pessoas estão em determinado local, e quem são, mas interessa sobretudo saber o que estão a fazer. O reconhecimento de actividades humanas é actualmente, uma das aplicações mais promissoras dos sistemas de vigilância.

A maior parte do trabalho desenvolvido, na área de reconhecimento de actividades humanas, tem-se focado no reconhecimento de comportamentos individuais, [1], [2] ou na detecção de actividades anómalas [3], [4]. Recentemente muitos autores começaram a estudar as interacções entre pessoas, [5], [6], [7]. Apesar destes autores já terem realizado avanços significativos na resolução do problema de reconhecimento de interacções humanas, este é ainda um problema em aberto e muito actual, uma vez que ainda não se sabe como caracterizar de forma eficaz a interacção entre pessoas, nem existem métodos de classificação capazes de operar de forma totalmente automática e robusta em ambientes reais.

1.2 Descrição do Problema

O problema que esta tese se propõe resolver, consiste no reconhecimento de interacções, entre duas pessoas com recurso a uma única camera de vídeo. Foram escolhidas quatro interacções que se pretende identificar: abraço, cumprimento, cruzamento e luta. Nesta escolha, procurou-se incluir exemplos de cada uma das classes de interacção humana, neutras (cruzamento), positivas (abraço e cumprimento) e negativas (luta). No caso da classe de interacções positivas foram escolhidas duas interacções, pois pretendia-se ter duas actividades que apresentassem algum grau de semelhança, para no fim se poder aferir sobre a sensibilidade do sistema proposto.

Uma vez que este problema é bastante complexo são realizadas algumas hipóteses com vista à simplificação do mesmo. Em relação ao ambiente onde são realizadas as interacções é assumido que o fundo é estático, que a iluminação é estável e que existem no máximo duas pessoas em cena. Em relação às pessoas que realizam as interacções, são actores que se deslocam perpendicularmente ao campo de visão da camera, interagem perto da camera mas com todo o corpo dentro do campo de visão da mesma. Admite-se ainda que as pessoas utilizam peças de roupa de cor diferente no tronco e nas pernas. Finalmente em relação à camera, a sua posição é estática, captura imagens a cores, e tem projecção perspectiva.

1.3 Metodologia

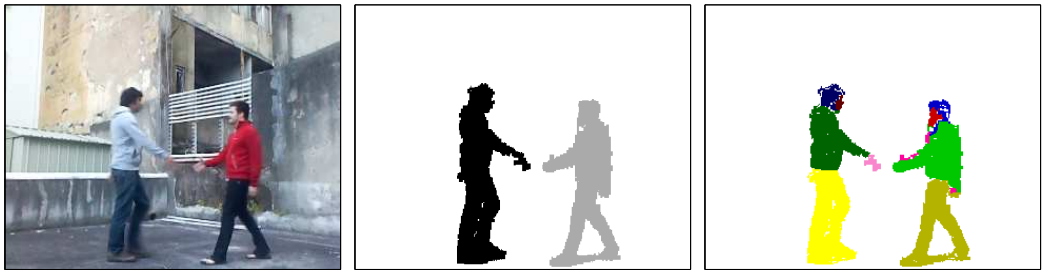
Para resolver o problema proposto nesta tese, vão ser utilizadas duas aproximações que diferem no modelo utilizado para representar o corpo humano, e, consequentemente, nas características extraídas para classificar as interacções.

A primeira aproximação utiliza um modelo da silhueta para representar o corpo humano. Ou seja, cada indivíduo é representado pela sua silhueta, que é segmentada e seguida em todos os frames de uma sequência de vídeo.

A segunda aproximação utiliza um modelo anatómico para representar o corpo humano. Este modelo é um refinamento do modelo da silhueta, que procura, à custa de passos de processamento adicionais, segmentar as diversas silhuetas detectadas em partes anatómicas coerentes.

A Figura 1.1 mostra um exemplo de um frame de entrada do sistema e dos dois modelos utilizados para representar o corpo humano. A Figura 1.2 mostra o diagrama de blocos

do sistema implementado nesta tese, constituído por quatro módulos de processamento: subtracção de fundo, seguimento de regiões activas, estimação de partes anatómicas e extracção de características. É de salientar a relação entre os dois modelos estudados, com o bloco de estimação de partes anatómicas a ser o responsável pelo refinamento do modelo da silhueta que permite obter o modelo anatómico.



(a) (b) (c)

Figura 1.1: Exemplo de uma imagem de entrada e dos dois modelos utilizados para representar o corpo humano: (a) imagem de entrada, (b) modelo da silhueta, (c) modelo anatómico.

1.4 Dados Experimentais

Os dados vídeo utilizados para testar o sistema foram adquiridos utilizando uma webcam Motion Eye 1.3MP embutida no computador da Sony, Vaio VGN-FW21M. Decidiu-se utilizar esta camera pois é um dispositivo comum, que equipa a maior parte dos computadores portáteis actuais, e que captura imagens com a qualidade semelhante à que se obtém nas câmaras correntes de video-vigilância. O computador foi assente numa base estável, com a webcam orientada de forma paralela ao chão. Capturaram-se interacções entre duas pessoas em ambiente exterior, ainda que com iluminação estável e com um fundo complexo. Os dados vídeo foram convertidos para o formato Windows AVI sem qualquer tipo de compressão, com 15 frames/s e uma resolução de 320x240 pixeis. A qualidade da cor é de 24 bit (isto é, 8 bit para cada canal de cor R,G e B).

Foram capturadas sequências das seguintes interacções: abraço, cruzamento, cumprimento e luta. Para cada uma das interacções foram utilizados 10 pares de pessoas, de entre um conjunto de 7 voluntários, perfazendo um total de 40 sequências (4 interacções $\times$ 10 pares de pessoas). Cada sequência contém apenas uma interacção entre duas pessoas e começa sempre sem nenhum indivíduo na imagem, com estes a partirem de posições afastadas aproximando-se a passo até interagirem. Os sujeitos usaram vários tipos de roupa casual, e foram instruídos para interagir uns com os outros de maneira natural nos quatro tipos de interacções. A duração das sequências varia entre os 2 e os 9 segundos dependendo da interacção e das pessoas envolvidas. A Tabela 1.1 indica o número total de imagens capturadas para cada uma das interacções.

A figura 1.3 ilustra os quatro tipos de interacções capturadas, em três momentos distintos de cada sequência.

1.5 Organização da Tese e Contribuições

O trabalho está estruturado da seguinte forma. No primeiro capítulo, de introdução, é descrito o problema as suas especificações e hipóteses, e são brevemente descritas as aprox-

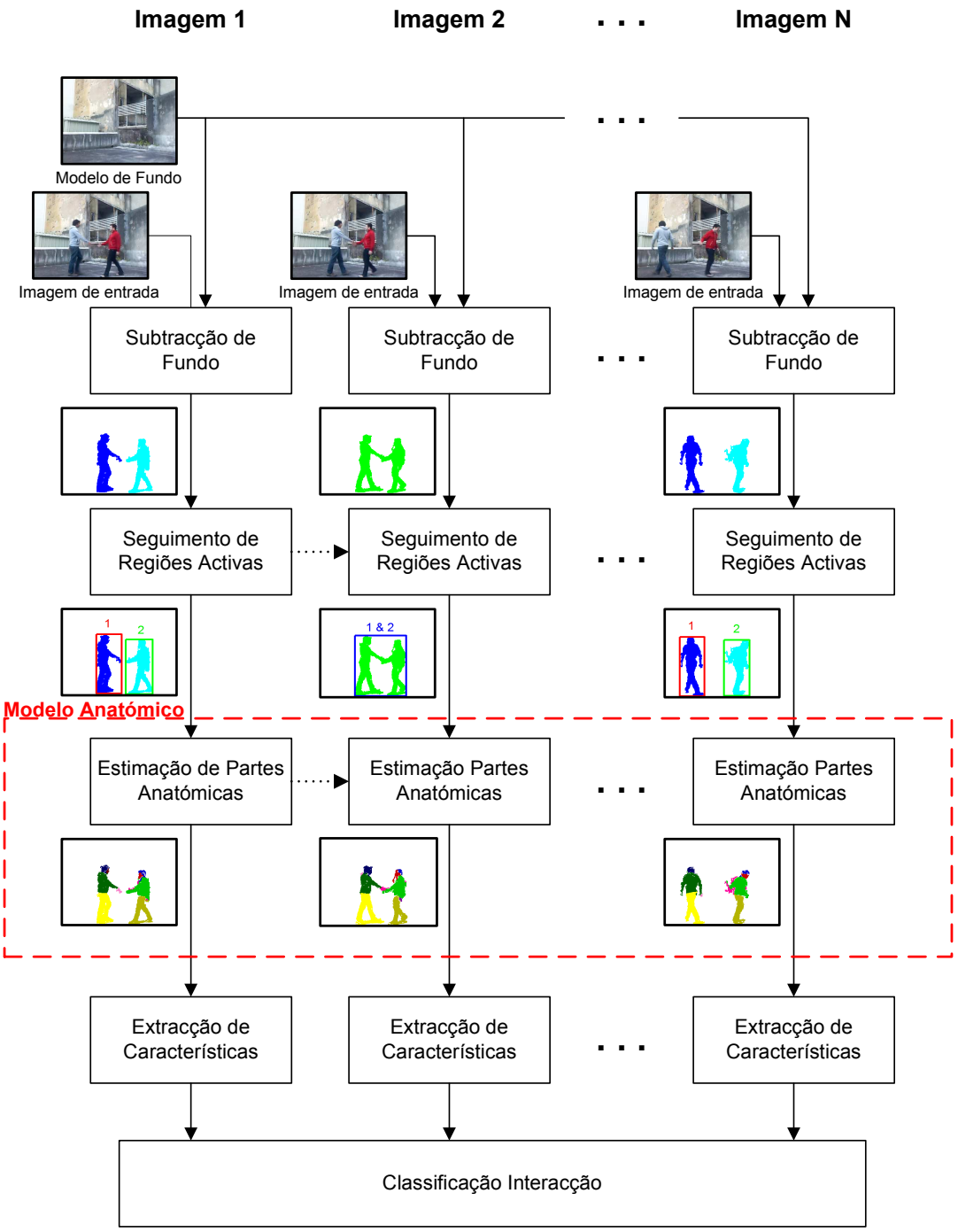


Figura 1.2: Diagrama de blocos do sistema proposto para identificar interacções.

Tabela 1.1: Número de imagens capturadas por actividade.

	Abraço	Cruzamento	Cumprimento	Luta	Total
Nº Imagens	936	755	482	735	3108

imações utilizadas na sua resolução. É ainda descrito o conjunto de dados experimentais adquiridos para testar o sistema implementado.

No Capítulo 2, o reconhecimento de interacções humanas é decomposto em três sub-problemas, e são apresentados os mais recentes avanços na resolução de cada um destes problemas.

No Capítulo 3 é apresentado o método utilizado para segmentar e seguir pessoas, que permite estimar e manter ambos os modelos de representação humana estudados. Primeiro descrevem-se todos os passos necessários na estimação do modelo da silhueta, e de seguida os passos adicionais necessários na estimação do modelo anatómico. São ainda apresentados os resultados de cada um dos módulos descritos nesta secção, bem como as conclusões sobre o desempenho dos mesmos.

No Capítulo 4 são descritos os módulos responsáveis pelo reconhecimento de actividades. É introduzido o classificador utilizado, e são descritas as características extraídas para cada modelo do corpo humano. Apresentam-se também os resultados da classificação e conclusões sobre os mesmos.

Por fim, nos Capítulos 5 e 6, de conclusões, apresenta-se uma avaliação global do sistema implementado e descrevem-se possíveis trabalhos futuros.

As principais contribuições desta tese para o problema de reconhecimento de interacções entre duas pessoas numa sequência vídeo são:

- Aquisição de um conjunto de sequências de video, que permitem avaliar o desempenho de um sistema de reconhecimento de interacções.
- Resolução do problema de oclusão entre pessoas quando é utilizado um modelo da silhueta.
- Segmentação de duas pessoas sobrepostas numa imagem quando é utilizado um modelo anatómico.
- Definição e extracção de características que permitem identificar interacções entre pessoas quando estas são representadas por um modelo da silhueta.
- Definição e extracção de características que permitem identificar interacções entre pessoas quando estas são representadas por um modelo anatómico.
- Arquitectura de um sistema de reconhecimento de interacções humanas.

Figura 1.3: Exemplos das quatro actividades: abraço (1ª linha), cumprimento (2ª linha), cruzamento (3ª linha) e luta (4ª linha).

Capítulo 2

Estado da Arte

Este capítulo resume o trabalho anterior sobre reconhecimento de actividades humanas, e identifica algumas das principais contribuições neste tema.

A identificação de actividades humanas pode ser decomposta em três sub-problemas [8]: segmentação do corpo humano a partir de imagens, seguimento do mesmo e reconhecimento da actividade. Adoptaremos esta divisão e descreveremos métodos existentes para cada um destes problemas.

2.1 Segmentação do Corpo Humano

Um vídeo contém uma quantidade enorme de informação, pois descreve a evolução de componentes de cor de cada pixel da imagem ao longo do tempo. No entanto a maioria desta informação não é relevante para a tarefa de identificação e reconhecimento de actividades. É necessário utilizar mecanismos que permitam detectar objectos de interesse, extrair a sua fronteira, e segui-la nas imagens seguintes.

Um algoritmo de detecção de objectos deve ter os seguinte requisitos [9], precisão na detecção dos contornos do objecto (precisão espacial) e estabilidade de detecção ao longo do tempo (coerência temporal). Deve ser capaz de detectar mudanças de baixa magnitude (sensibilidade) e fornecer resultados precisos quando as condições são variáveis, tais como variações de iluminação (robustez).

Apesar de ser um problema muito estudado, a detecção de objectos em ambientes complexos é um problema em aberto [10], [11]. Apresentam-se de seguida alguns dos métodos de detecção de objectos mais populares.

2.1.1 Fluxo Óptico

O fluxo óptico é uma imagem vectorial que representa o movimento aparente de cada elemento da imagem, entre duas imagens consecutivas. O fluxo óptico fornece informações precisas, tanto sobre as regiões que se encontram em movimento, como sobre a direcção e magnitude do movimento.

A maioria dos métodos utilizados para calcular o fluxo óptico assumem que a cor de pixeis homólogos em duas imagens consecutivas não varia (constância de cor). Na prática o cálculo do fluxo óptico é afectada por vários erros, causados por variação de iluminação, por ruído de aquisição, por falta de textura na vizinhança do pixel ou por oclusões. O fluxo óptico foi utilizado em vários trabalhos de análise de movimento humano p.ex., [12], [13]. No entanto, as dificuldades mencionadas e a falta de robustez têm impedido uma

utilização mais generalizada do fluxo óptico em vigilância. Uma revisão sobre este tema pode ser encontrada em [14].

2.1.2 Subtracção de Fundo

Se a camera estiver numa posição estática é possível construir uma imagem de fundo que corresponde a uma situação em que não há objectos móveis na cena. A detecção de objectos móveis pode ser obtida comparando cada imagem adquirida com a imagem de fundo e detectando desvios significativos. Qualquer mudança significativa numa região da imagem em relação ao fundo significa a existência de um objecto. Normalmente, os pixeis detectados como activos são marcados para processamento posterior e sujeitos a um algoritmo de componentes conexas, para formar regiões conexas que correspondem a objectos. Este processo é chamado de subtracção de fundo.

O algoritmo de subtracção de fundo básico tem sido melhorado de várias formas. Uma das melhorias foi proposta em [15] onde se considera cada pixel como uma variável aleatória com distribuição normal no espaço de cor YUV. Apesar de hoje em dia, existirem na literatura inúmeros algoritmos de subtracção de fundo, a maioria deles segue um simples diagrama de fluxo, definido em [16], composto por quatro passos, que são (1) pré-processamento (tarefas simples de processamento de imagem que transformam o vídeo de input original para um formato que pode ser processado pelos passos seguintes), (2) modelação do fundo (também conhecido como manutenção do fundo), (3) detecção de regiões activas (também conhecido como subtracção de fundo) e (4) validação dos dados (também referido por pós-processamento, utilizado para eliminar pixeis que não correspondem a objectos).

2.1.3 Respostas de Filtros

Um outro caminho alternativo consiste na aplicação de filtros espaço-temporais [3]. A sequência de imagens é processada utilizando uma Gaussiana espacial e uma derivada de Gaussiana ao longo do eixo temporal. Devido à derivação temporal, o filtro apresenta respostas com intensidades altas em zonas de movimento. Esta resposta é depois comparada com um limiar pré-fixado para formar uma máscara binária que é depois agregada num histograma espacial. Este histograma espacial contém informação sobre o movimento e a sua localização espacial de uma forma compacta e é bastante útil em vídeo vigilância distante.

A noção de filtragem espaço-temporal é utilizada também por outros investigadores [17], onde se propõe uma generalização do detector de cantos de Harris para utilização em vídeos, recorrendo a um conjunto de filtros de derivadas Gaussianas espaço-temporais. Ainda outra aproximação é apresentada em [18] onde são extraídos pontos de referência periódicos, baseados no movimento, utilizando uma Gaussiana no espaço e uma função de Gabor ao longo do tempo. Uma vez que estas aproximações são baseadas em operações simples de convolução, são fáceis de implementar e de rápida computação. Este tipo de detecção é bastante útil em cenários onde a resolução é baixa, ou a qualidade do vídeo é pobre, impossibilitando a aplicação de outras técnicas como o fluxo óptico ou a subtracção de fundo.

2.2 Seguimento de Objectos

O objectivo de um algoritmo de seguimento de objectos é gerar a trajectória de um objecto ao longo do tempo, localizando a sua posição em cada imagem de uma sequência de vídeo. A tarefa de detectar o objecto e estabelecer a correspondência temporal entre as localizações do objecto ao longo das várias imagens, pode ser realizada em conjunto ou separadamente. No primeiro caso, a região que o objecto ocupa e a sua correspondência são efectuadas em conjunto actualizando iterativamente a localização e a região do objecto, obtida em imagens anteriores. No segundo caso, são geradas, por um algoritmo de detecção de objectos, regiões passíveis de conter o objecto, realizando-se de seguida a tarefa de correspondência ao longos das vários imagens. Em ambos os casos, o objecto é representado utilizando um modelo de aparência, a Figura 2.1 ilustra vários modelos de aparência. O modelo utilizado para representar o objecto, limita o tipo de movimento e deformação a que o objecto pode estar sujeito. Se for escolhido representar o objecto como um ponto, só se pode descrever a trajectória do mesmo ao longo do tempo como uma translação. No caso de ser utilizado uma forma simples para representar o objecto (p.ex., elipse), modelos de movimento paramétricos como transformações afins e projectivas já são apropriadas. Estas representações podem aproximar o movimento de objecto rígidos presentes na cena, no entanto para objectos articulados, a silhueta e o contorno são representações que descrevem melhor o objecto e possibilitam a utilização de modelos tanto paramétricos como não paramétricos para descrever o movimento do objecto. A Figura 2.2 mostra alguns modelos utilizados na representação do corpo humano e o seu relacionamento com os diversos métodos de seguimento. Em seguida descrevem-se alguns métodos de seguimento.

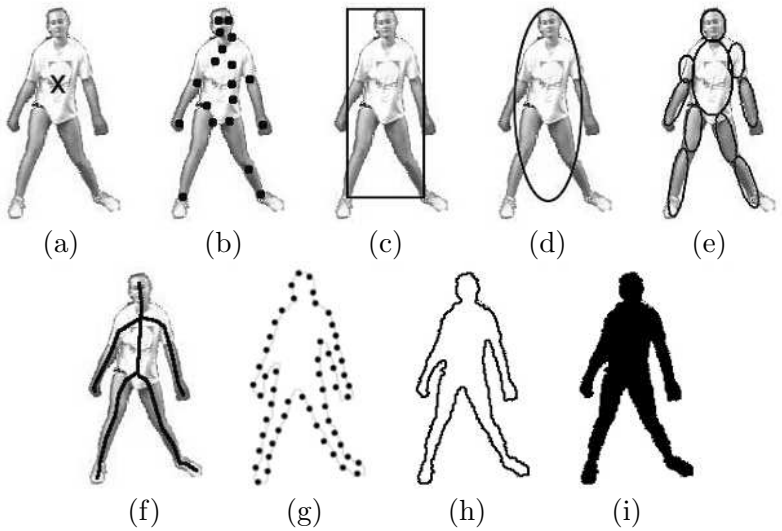


Figura 2.1: Representações objecto. (a) Centroide, (b) múltiplos pontos, (c) rectângulo, (d) elipse, (e) múltiplas elipses com base nas várias partes do objecto, (f) esqueleto objecto, (g) pontos de controlo no contorno do objecto, (h) contorno do objecto, (i) silhueta do objecto. Figura adaptada de [9].

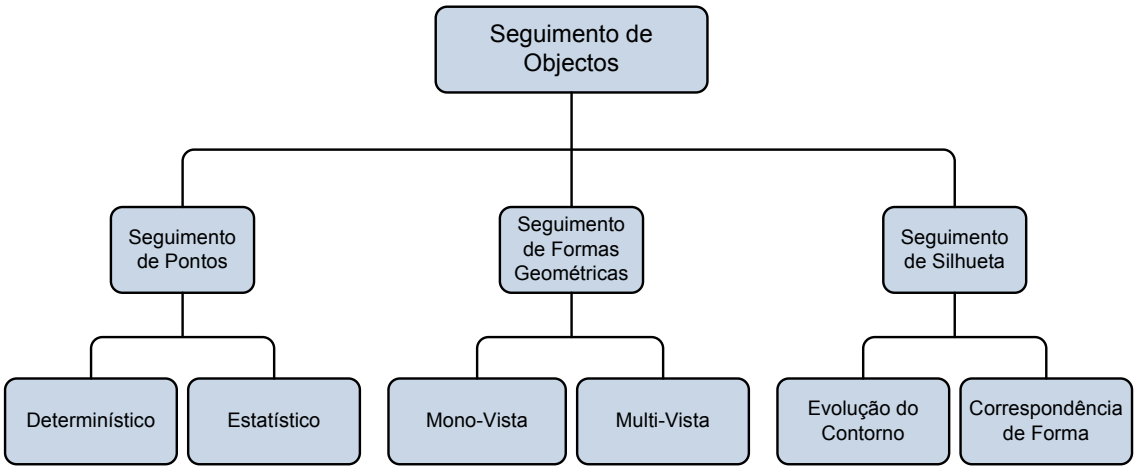


Figura 2.2: Taxonomia dos métodos de seguimento. Figura adaptada de [9].

2.2.1 Seguimento de Pontos

Os objectos detectados em cada imagem são representados como um ponto (p.ex., centro de massa), e a associação dos pontos em imagens consecutivas é realizada com base no movimento e posição prévia do objecto. A correspondência entre pontos é um problema difícil especialmente na presença de oclusões, falsas detecções, entradas e saídas de cena dos objectos.

Em geral os métodos de seguimento de pontos podem dividir-se em duas categorias, métodos determínisticos e métodos estatísticos. Os métodos determínisticos definem um custo para associar cada objecto na imagem $t-1$ a um objecto na imagem t utilizando um conjunto de restrições de movimento. Estas restrições podem ser, distância, variações de velocidade, entre outras. O problema é então formulado como um problema de associação entre os objectos detectados no instante $t-1$ e no instante t. Os métodos estatísticos utilizam não só medidas do objecto, mas têm também em conta incertezas, no movimento e no processo de medição da posição do objecto para realizar a correspondência. É utilizado um modelo dinâmico para modelar a evolução das propriedades do objecto como a posição, velocidade e aceleração que tem em conta as incertezas. A associação pode ser feita usando algoritmos de optimização [19], filtros de associação de dados [20], *Multiple Hypotheses Tracking* [21] ou filtros de partículas [22].

2.2.2 Seguimento com base em Kernel

O seguimento com base em *kernel* é tipicamente realizado calculando o movimento do objecto, que é representado por uma forma geométrica primitiva, de uma imagem para a outra. O movimento do objecto é normalmente descrito sob a forma de uma transformação geométrica (translação, afim, etc.) ou sob a forma de um mapa de fluxos.

Os algoritmos de seguimento com base em *kernel* diferem em termos da representação do objecto utilizada, número de objectos seguidos e métodos utilizados para calcular o movimento. Pode-se então dividir estes métodos em duas categorias, modelos de representação mono-vista ou multi-vista. Nos modelos mono-vista podem ser empregues dois tipos de aproximações, correspondência com base em um *template*, [23][24], normalmente composto ou por intensidades de imagem, ou características de cor dos pixeis dentro da forma geométrica, ou histogramas de cor, ou misturas de Gaussianas dentro da mesma.

Neste tipo de correspondência é efectuada uma busca à procura de uma região na imagem $t+1$ semelhante ao *template* definido em t. A outra aproximação, [25][26] recorre ao fluxo óptico para calcular a translação dos pixeis dentro da forma geométrica primitiva. Em ambos os métodos, os modelos de aparência, são normalmente gerados *online*, logo representam informação de observações recentes do objecto. Se o objecto mudar de pose drasticamente durante o processo de seguimento, o modelo de aparência pode já não ser válido. É com vista à resolução deste problema que surgem os métodos multi-vista, em que um modelo de aparência multi-vista é gerado *offline*, sendo depois utilizado durante o processo de seguimento. O modelo de aparência pode ser, como proposto em [27] construído com recurso à Análise de Componente Principal, sendo depois a correspondência entre região da imagem que contém o objecto e modelo efectuada com recurso à minimização da diferença entre os valores próprios do modelo e a imagem. Ou outra aproximação proposta em [28] consiste em utilizar várias vistas do objecto para treinar uma máquina de vectores de suporte, que é depois utilizada para classificar regiões da imagem ou em objecto ou em fundo.

2.2.3 Seguimento de Silhueta

Os objectos que estão a ser alvos do processo de seguimento, podem ter formas complexas, por exemplo, mãos, cabeças, pernas, que não são passíveis de serem correctamente descritas por formas geométricas simples. Os métodos de seguimento com base em silhuetas fornecem um método de descrição de formas capaz de caracterizar este tipo de objectos. Podemos dividir os métodos de seguimento com base em silhuetas em duas categorias, nomeadamente, correspondência de forma e evolução do contorno.

A aproximação baseada na correspondência de forma, [29][30][31], pode ser realizada à semelhança do seguimento de *template* (Secção 2.2.2), realizando a correspondência entre a silhueta de um objecto e uma região da imagem em análise. O modelo do objecto, que é usualmente um mapa de contornos, é reinicializado, por forma a lidar com as mudanças de aparência, em todas as imagens depois de o objecto ser detectado.

Os métodos de seguimento com base na evolução de contorno, reajustam iterativamente um contorno inicial na imagem $t-1$, para a sua nova posição na imagem t. Este tipo de métodos requer que uma parte do objecto na imagem actual se sobreponha à região que o objecto ocupava na imagem anterior. Existem dois tipos de aproximações utilizadas para implementar este método. A primeira aproximação, [32][33], consiste em utilizar modelos de espaço de estados, para modelar a forma do contorno e o seu movimento. O estado do objecto é actualizado a cada instante de tempo por forma a maximizar a probabilidade *a posteriori* do contorno. A probabilidade *a posteriori* depende do estado anterior e da verosimilhança actual que é usualmente definida em termos da distância do contorno à fronteira observada. A segunda aproximação, [34][35][36], evolui directamente o contorno minimizado a energia do contorno com recurso a técnicas de minimização como o gradiente descendente. A energia do contorno é definida em termos de informação temporal (fluxo óptico) [34][35], ou como estatísticas de aparência geradas a partir do objecto e das regiões de fundo [36].

2.3 Reconhecimento de Actividades

O reconhecimento de actividades encontra-se ainda longe do seu máximo potencial. Nos últimos anos, fruto dos novos desenvolvimentos tecnológicos e do aumento de interesse público por esta área, tem-se assistido a um aumento na investigação. Cada vez mais

autores estão interessados em contribuir para os avanços nesta área, o que se reflecte no grande número de diferentes ideias e aproximações ao problema existentes hoje em dia. As aproximações actuais podem no entanto ser divididas em três grandes classes [37]: não paramétricas, volumétricas e paramétricas. As aproximações não paramétricas tipicamente extraem um conjunto de características de cada imagem do vídeo. As aproximações volumétricas, por outro lado, consideram o vídeo como sendo um volume 3D de intensidades dos pixeis, para a partir daí extrair características. As aproximações paramétricas impõem um modelo na dinâmica temporal das actividades, cujos parâmetros são estimados a partir de dados de treino. A Figura 2.3 apresenta uma taxonomia dos diferentes métodos de reconhecimento de actividades. Descrevem-se de seguida alguns dos métodos mais populares de reconhecimento de actividades.

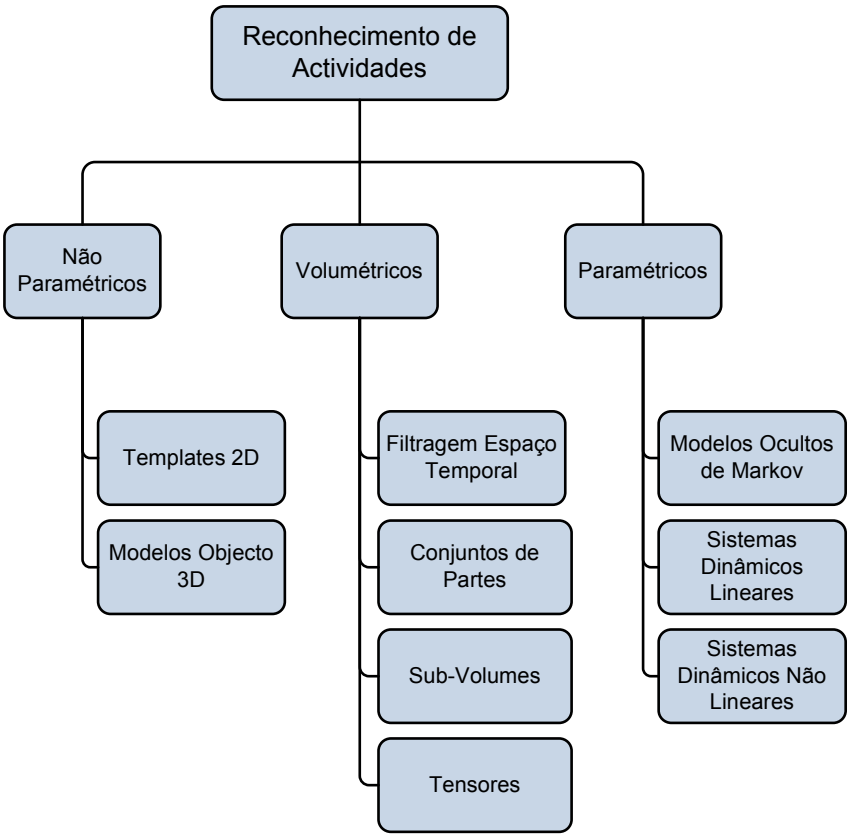


Figura 2.3: Taxonomia dos métodos de reconhecimento de actividades. Figura adaptada de [37].

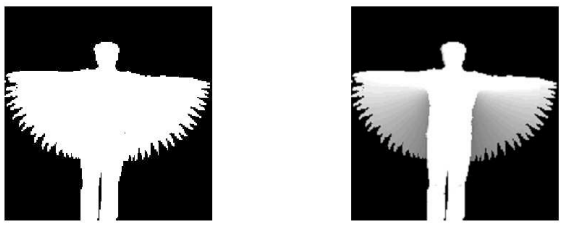
2.3.1 Aproximações Não Paramétricas

Nas aproximações não paramétricas os vídeos analisados são tratados como uma sequência de imagens, onde é necessário segmentar e seguir as pessoas presentes em cena. As aproximações não paramétricas podem dividir-se em duas categorias, *templates* 2D e modelos objecto 3D.

Os métodos que utilizam *templates* 2D agregam uma sequência de regiões activas segmentadas numa única imagem estática. Em [38] são propostos dois métodos de agregação, no primeiro método todas as imagens segmentadas têm o mesmo peso o que leva à representação *Motion Energy Image* (MEI). O segundo método atribui um peso maior às

imagens mais recentes da sequência, e um peso menor às imagens mais antigas, o que leva a uma representação chamada *Motion History Image* (MHI). A Figura 2.4 mostra um exemplo destas representações. O MEI e o MHI juntos são considerados um *template* de uma determinada actividade. A partir do *template*, translações, rotações e momentos de Hu [39] são extraídos e usados no reconhecimento. O MEI e o MHI juntos têm poder discriminativo suficiente para reconhecer actividades simples, no entanto para acções mais complexas perdem algum do seu poder devido a se sobre-escrever a história do movimento [38].

Nos métodos em que se usam modelos do objecto 3D as actividades são representadas como objectos espaço-temporais. Esta representação é estimada, empilhando uma sequência de silhuetas 2D ao longo do tempo. Uma sequência de silhuetas 2D (x, y) é tratada como um objecto no espaço (x, y, t). A Figura 2.5 ilustra um exemplo desta representação. A partir desta representação é proposto em [2] extrair descritores da superfície do objecto que correspondem a características geométricas como, picos, vales e pontos de sela. Alternativamente em [40] é proposto extrair descritores do objecto 3D resolvendo uma equação de Poisson. Estas aproximações requerem algoritmos de segmentação de fundo muito precisos, estando por isso limitadas a aplicações onde a camera é estática.



(a) (b)

Figura 2.4: Templates temporais, MEI (a) e MHI (b) obtidos agregando uma sequência de silhuetas 2D. Figura adaptada de [38].

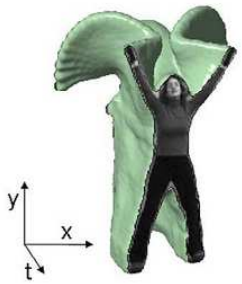


Figura 2.5: Objecto espaço temporal 3D, obtido empilhando uma sequência de silhuetas 2D. Figura adaptada de [40].

2.3.2 Aproximações Volumétricas

Nas aproximações volumétricas o vídeo em vez de ser considerado como uma sequência de imagens é tratado como sendo um volume 3D de intensidades dos pixeis. As aproximações volumétricas podem ser divididas em quatro classes, filtragem espaço-temporal, conjuntos de partes, sub-volumes e tensores.

As aproximações que recorrem a filtragens espaço-temporais utilizam a resposta de um banco de filtros quando aplicado a um vídeo para obter características que permitam reconhecer actividades. Por exemplo em [41], é utilizado um banco de filtros de Gabor com várias orientações e escalas espaciais e uma única escala temporal, para estimar modelos de aparência local para cada pixel. Uma determinada actividade é identificada utilizando a probabilidade espacial de cada pixel. Como os vídeos são analisados utilizando apenas uma escala temporal, este método não é aplicável a actividades que variem no seu tempo de execução. Uma extensão desta aproximação que tenta resolver este problema foi proposta em [42] onde são extraídos histogramas locais de gradientes espaço-temporais normalizados. A soma chi-quadrado entre cada histograma é então utilizada para realizar a correspondência com um *template*. As aproximações de filtragem espaço-temporal são rápidas e fáceis de implementar, devido a utilizarem apenas operações de convolução. No entanto como na maior parte das aplicações a largura de banda apropriada não é conhecida *a priori*, é necessário utilizar um banco de filtros de grande dimensão com várias escalas espaciais, para conseguir efectivamente descrever a dinâmica das actividades. Tal facto torna-se proibitivo pois a resposta de cada filtro é da mesma dimensão do volume de entrada, o que implica elevados custos computacionais.

As aproximações que se baseiam em conjuntos de partes consideram o volume de vídeo como um conjunto de partes locais, onde cada parte consiste num padrão de movimento distinto. Em [43] é proposto representar as acções com um modelo saco-palavras. O modelo saco-palavras é estimado utilizando uma generalização espaço-temporal do detector de Harris, para extrair pontos de interesse espaço-temporais. Estes pontos de interesse, podem então ser utilizados em conjunto com métodos de aprendizagem automática, como máquinas de vectores de suporte [44], para reconhecer actividades. O facto de se utilizar um modelo saco-palavras faz com que a geometria global entre partes locais não seja modelada. Actividades diferentes podem ser compostas por partes espaço-temporais semelhantes, mas a relação geométrica entre as diversas partes ser completamente diferente. A integração de geometria global na representação de vídeos por partes foi investigada em [45], [46]. Estas aproximações utilizam um modelo denominado constelação de partes. O custo computacional de se utilizar um modelo de constelação de partes pode ser muito elevado quando uma actividade é composta por muitas partes, como é o caso das actividades humanas. Uma aproximação para a minimização deste problema é apresentada em [47], onde é proposto utilizar um modelo hierárquico onde o nível mais alto é uma constelação de partes, composta por muito menos partes do que as partes que constituem o vídeo. O nível mais baixo é constituído por um saco-palavras, ou seja cada parte da constelação consiste num saco-palavras. Esta aproximação combina as vantagens da utilização dos dois modelos e é eficiente do ponto de vista computacional. Na maior parte destas aproximações a detecção de partes é usualmente conseguida utilizando operações lineares como, filtragens e gradientes espaço-temporais, logo é sensível a ruído, oclusões, mudanças de aparência, etc. Também já foi notado, que em actividades humanas em que os movimentos são suaves, os pontos de interesse detectados são muito escassos [43].

As aproximações que utilizam sub-volumes diferenciam-se das aproximações com base em conjunto de partes, pois é realizada uma correspondência directa entre sub-volumes que compõem um volume vídeo e *templates* armazenados. Alguns exemplos de utilizações deste método foram propostos em [48], [49] e [50], onde os sub-volumes são obtidos agrupando pixeis com base na sua aparência e localização espacial. Um *template* de uma actividade é depois correspondido, procurando pelo conjunto mínimo de sub-volumes que maximizam a sobreposição entre a sua união e o *template*. As aproximações com base em correspondência de sub-volumes são sensíveis a mudanças no fundo, no entanto são bastante robustas a

ruído e oclusões.

Nas aproximações em que se utilizam tensores, o volume espaço tempo 3D é considerado um tensor com três dimensões independentes. Em [51] é proposto modelar actividades humanas e a identidade das pessoas como dimensões independentes de um tensor. Realizando a decomposição do tensor que representa o volume 3D correspondente ao vídeo em modos dominantes (utilizando uma generalização da análise de componentes principais) é possível reconhecer actividades e identificar as pessoas que as executam. Outras aproximações semelhantes, que consideram um volume vídeo um tensor tridimensional foram propostas em [52], [53]. As aproximações com base em tensores oferecem uma grande flexibilidade, pois permitem incorporar características como o fluxo óptico ou respostas de filtros simplesmente adicionando mais uma dimensão ao tensor.

2.3.3 Aproximações Paramétricas

As aproximações paramétricas modelam as actividades como uma sequência estocástica de acções, ou seja, as actividades humanas são descritas num espaço de estados, onde a evolução temporal é modelada como uma sequência de saltos probabilísticos de um estado para outro. As aproximações paramétricas podem dividir-se em três classes, modelos ocultos de Markov (HMM), sistemas dinâmicos lineares (SDL) e sistemas dinâmicos não lineares (SDNL).

Uma das primeiras aplicações de modelos ocultos de Markov ao reconhecimento de actividades foi proposta em [54] onde se pretendia reconhecer vários tipos de pancadas de ténis como, *backhand, forehand, volley, smash,* etc modelando uma sequência de regiões activas segmentadas como saídas de um HMM. A partir desta aplicação muitos investigadores utilizaram HMM's para modelar actividades simples [55], [56], [57]. Estas aproximações são todas baseadas no pressuposto que apenas um indivíduo realiza a actividade, não tendo aplicação efectiva quando vários pessoas realizam uma actividade conjunta ou interagem entre elas. Este problema foi abordado em [58], onde é proposto modelar as interacções com conjuntos acoplados de HMM's. A partir daí mais autores realizaram avanços na modelação e reconhecimento de actividades utilizando HMM's, [59], [60], [61], [62]. Os HMM's são eficientes na modelação de sequências temporais, e são úteis pela sua capacidade discriminativa. No entanto devido aos pressupostos da dinâmica Markoviana e à natureza da invariância temporal intrínseca ao modelo, a utilidade dos HMM´s restringe-se a actividades relativamente simples e que sejam invariantes no tempo.

Os sistemas dinâmicos lineares são uma forma mais generalizada de HMM's, com o espaço de estados a poder assumir valores contínuos em $\mathbb{R}^k$ (k é a dimensão do espaço de estados), ao invés de estar restrito a uma série finita de símbolos. Os SDL's podem ser interpretados como uma generalização dos HMM´s a um espaço de estados contínuo com um modelo de observação Gaussiano. Existem inúmeras aplicações de reconhecimento de actividades com recurso a SDL's, [63], [64], [65]. Recentemente novos métodos de aprendizagem dos parâmetros intrínsecos do modelo a partir de dados de treino, [66], [67], tornaram os SDL's mais populares no reconhecimento de actividades, e muito mais fáceis de treinar que os HMM's. No entanto tal como os HMM's, os SDL's também são baseados nos pressupostos da dinâmica Markoviana e não se adequam ao reconhecimento de actividades complexas.

Os sistemas dinâmicos não lineares surgem como uma resposta às restrições dos HMM's e SDL's a actividades lineares e de dinâmica estacionária. Uma actividade complexa como uma pessoa pegar num objecto, caminhar até uma mesa, pousar o objecto e finalmente sentar-se, pode ser vista como uma sequência de pequenos segmentos em que cada um pode

ser modelado usando um SDL. O processo inteiro pode então ser visto como uma troca de SDL's. Existem hoje em dia vários autores a utilizar esta aproximação ao problema, por exemplo, [68], [69], [70] em que cada um propõe uma aproximação diferente para modelar as trocas. Os SDNL's têm uma capacidade descritiva e de modelação muito maior que os HMM's e os SDL's, contudo o processo de treino é imensamente mais complicado. Na prática os SDNL's requerem uma enorme quantidade de dados de treino, ou então afinação manual extensiva.

Capítulo 3

Seguimento de Pessoas

Este capítulo descreve os métodos de segmentação e seguimento de pessoas utilizados para construir as representações do corpo humano utilizadas nesta tese.

3.1 Introdução

A tarefa de reconhecimento depende da capacidade de ter algoritmos robustos, capazes de segmentar e seguir o corpo humano em movimento. Quando se utiliza o modelo da silhueta estes algoritmos são suficientes para proceder à extracção de características que permitam identificar as interacções. Quando se utiliza o modelo anatómico é preciso proceder a passos de processamento adicionais, que permitam segmentar o corpo humano em partes do corpo coerentes. Esta tarefa apresenta várias dificuldades, sobretudo quando há mais do que um indivíduo em cena, uma vez que é necessário distinguir a quem pertencem as diversas partes do corpo segmentadas e lidar com as oclusões entre indivíduos.

Com vista à resolução deste problema foi implementado um método baseado no trabalho apresentado por Aggarwal et al [71], onde é utilizada uma aproximação *bottom-up*, no sentido em que pixeis individuais são agrupados em regiões homogéneas e depois em partes do corpo. As diversas regiões detectadas são seguidas de forma automática ao longo de toda a sequência. Finalmente é utilizado conhecimento sobre o corpo humano para associar as diversas regiões a partes do corpo.

Segue-se uma descrição detalhada de cada módulo que compõe o sistema. A Figura 3.1 apresenta um diagrama de blocos do sistema proposto, que completa o sistema apresentado na Figura 1.2.

3.2 Subtracção de Fundo

A detecção de regiões activas e a sua segmentação é um passo essencial na solução do problema proposto. Para realizar esta tarefa é utilizado um método de subtracção de fundo proposto em [31].

Seja V uma imagem de cor no espaço de cor HSV. Designamos por $v(x, y)$

$$v(x, y) = [v_H(x, y), v_S(x, y), v_V(x, y)]^T. \tag{3.1}$$

o valor da imagem no ponto de coordenadas (x, y). Admite-se que $v(x, y)$ é uma variável aleatória com distribuição normal.

A média $\mu_Z(x, y)$ e o desvio padrão $\sigma_Z(x, y)$ da intensidade de cada pixel é calculada para cada canal de cor $Z \in \{H, S, V\}$, usando 20 imagens de treino, quando não se encontra

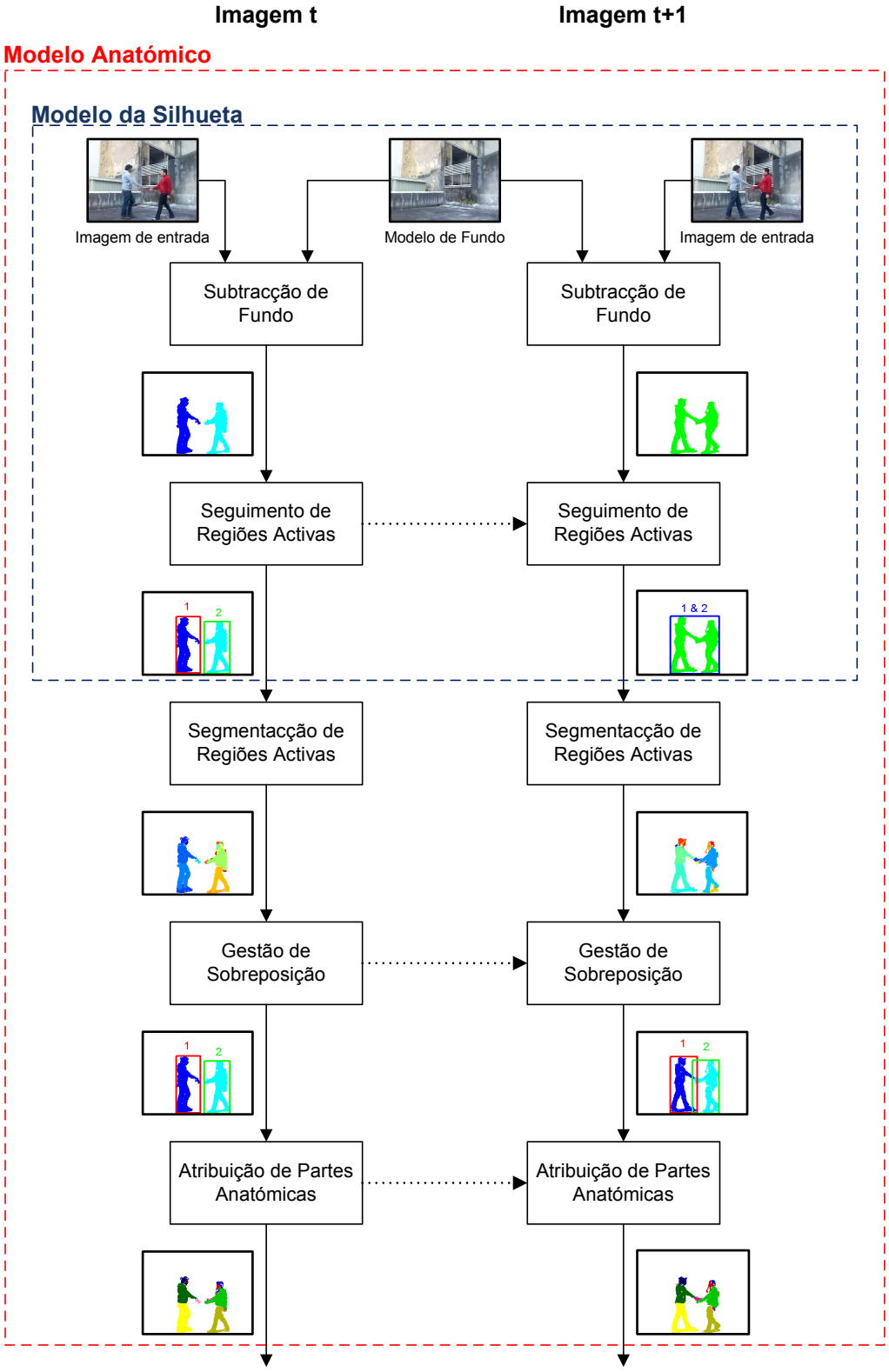


Figura 3.1: Diagrama de blocos do sistema proposto para segmentar e seguir as diversas partes do corpo humano.

nenhuma pessoa na imagem. Foram utilizados 20 imagens de treino, pois de um ponto de vista estatístico esse é considerado o número mínimo de amostras necessárias para calcular com fiabilidade a média e a covariância.

A segregação das regiões activas é conseguida avaliando a mudança de intensidade de cada pixel (x, y) numa determinada imagem de entrada, calculando a distância de Mahalanobis entre esta e o modelo do fundo $\delta_Z(x, y)$ para cada canal de cor Z.

$$\delta_Z(x, y) = \frac{|v_Z(x, y) - \mu_Z(x, y)|}{\sigma_Z(x, y)}. \tag{3.2}$$

A imagem contendo as regiões activas $F(x, y)$ é definida pelo máximo das três medidas de distância, δ_H, δ_S e δ_V correspondentes aos canais de cor H, S e V.

$$F(x, y) = max[\delta_H(x, y), \delta_S(x, y), \delta_V(x, y)]. \tag{3.3}$$

Cada pixel (x, y) da imagem F é de seguida comparado com um limiar para criar uma máscara binária. O valor do limiar utilizado foi determinado experimentalmente realizando uma série de testes em que comparou as regiões activas obtidas com diferentes limiares. No geral, limiares mais baixos produzem regiões activas maiores e mais ruído, enquanto limiares altos produzem regiões activas mais pequenas, possivelmente com buracos e menos ruído.

Após a subtracção do fundo, são realizadas operações morfológicas para eliminar regiões pequenas de ruído. A Figura 3.2 mostra um exemplo de uma imagem de entrada e as regiões activas detectadas como resultado da subtracção de fundo.

(a) (b)

Figura 3.2: Exemplo de uma imagem de entrada (a) e as regiões activas detectadas (b).

3.3 Seguimento de Regiões Activas

O seguimento de região activas ao longo de uma sequência vídeo é de extrema importância no contexto do problema proposto. É necessário conseguir manter a identidade de cada região activa em todas as imagens da sequência. Duas regiões activas em $t-1$ associadas a um indivíduo isolado, podem juntar-se em t numa uma única região activa em t que contém os dois indivíduos sobrepostos. É então necessário conseguir detectar não só junções mas também divisões. No caso de ser detectada uma divisão é ainda necessário conseguir estabeler correctamente a identidade de cada região activa. Este módulo tem então três

funções: estabelecer a correspondência entre regiões activas de imagens sucessivas, detectar junções e divisões e estabelecer a identidade de cada região activa após uma oclusão. A Figura 3.3 apresenta quatro momentos distintos de uma sequência de vídeo, com a primeira linha a corresponder à entrada do módulo de seguimento de regiões activas e a segunda linha à sua saída. Estes quatro momentos da sequência permitem ilustrar as três funções deste módulo, correspondência entre regiões activas da imagem #5 para a #25, detecção de junções da imagem #25 para a #40 e detecção de divisões e estabelecimento da identidades de cada região activa após uma oclusão, da imagem #40 para a #80.

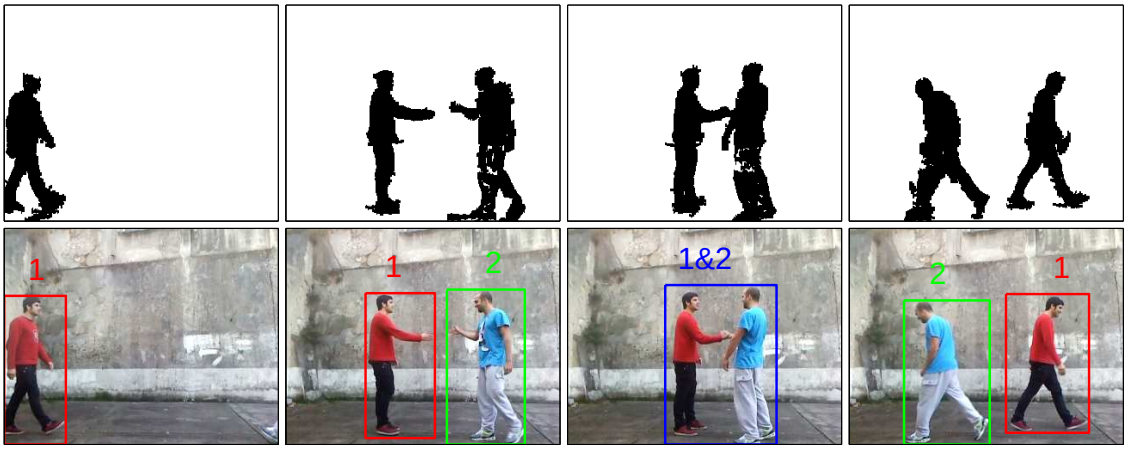


Figura 3.3: Exemplo da entrada do módulo de seguimento de regiões activas (1ºlinha) e a sua saída.

3.3.1 Correspondência Regiões Activas

O método de seguimento utilizado baseia-se numa aproximação proposta por Yang et al [72]. O seguimento é feito de forma incremental actualizando em cada imagem as trajectórias dos objectos activos na cena. Isso é feito associando as regiões activas detectadas no instante t (medidas) às regiões activas detectadas em $t-1$ (rastos). Vamos denominar os conjuntos de rastos e medidas existentes no instante t por R^t e M^t respectivamente. O processo começa com a construção de uma matriz de distâncias $\Delta^{t,t-1}$ entre os rastos R^t e as medidas M^t. A matriz de distâncias $\Delta^{t,t-1}$ é baseada na distância Euclideana,

$$\Delta_{ij}^{t,t-1} = \sqrt{(R_i^t x - M_j^t x)^2 + (R_i^t y - M_j^t y)^2} \tag{3.4}$$

Onde $R_i^t x$, $M_j^t x$, $R_i^t y$ e $M_j^t y$ representam o centroide das regiões activas R_i^t e M_j^t, $i = 1, ..., m$, $j = 1, ..., n$.

Cada elemento da matriz $\Delta^{t,t-1}$ calculado é comparado com um determinado limiar, se for maior é lhe atribuído o valor infinito. Com base na análise da matriz $\Delta^{t,t-1}$ é construída uma matriz de correspondências C^t com a mesma dimensão, que associa as medidas e os rastos. Apresentam-se de seguida os detalhes da sua construção.

1. Inicializar valores da matriz C^t a zero.

2. Encontrar o mínimo em todas as linhas $\alpha = \{\alpha_1, ..., \alpha_m\}$ e colunas $\beta = \{\beta_1, ..., \beta_m\}$ da matriz $\Delta^{t,t-1}$.

$$\Delta_{i,\alpha_i}^{t,t-1} = min(\Delta_{ij}^{t,t-1}), \quad j = 1, ..., n \tag{3.5}$$

$$\Delta^{t,t-1}_{\beta_j,j} = min(\Delta^{t,t-1}_{ij}), \quad i = 1, ..., m \tag{3.6}$$

3. Adicionar um a cada elemento correspondente da matriz C^t.

$$C^t_{i,\alpha_i} = C^t_{i,\alpha_i} + 1, \quad i = 1, ..., m \tag{3.7}$$

$$C^t_{\beta_j,j} = C^t_{\beta_j,j} + 1, \quad j = 1, ..., n \tag{3.8}$$

Existem três valores possíveis para os elementos da matriz C^t: zero, um e dois. Zero significa que não existe selecção. Um representa a existência de uma selecção. Dois significa que o rasto e a medida se seleccionam mutuamente. Podem existir cinco resultados possíveis na análise da matriz C^t:

- O rasto não está associada a nenhuma medida. (Todos os elementos da linha são zero)
- A medida não está associada a nenhum rasto. (Todos os elementos da coluna são zero)
- Uma medida está associada a um rasto. (Elemento da matriz igual a dois)
- Uma medida está associada a mais do que um rasto. (Mais do que um elemento da coluna é maior que zero)
- Um rasto está associado a mais do que uma medida. (Mais do que um elemento da linha é maior que zero)

Se um elemento da matriz C^t for igual a dois, a medida e o rasto são associados, e todos os elementos na mesma coluna e linha da matriz $\Delta^{t,t-1}$ são actualizados para infinito. Depois disso, uma nova matriz C^t é construída com base na matriz $\Delta^{t,t-1}$ actualizada. Este processo continua a repetir-se até nenhum dos elementos da matriz C^t ser igual a dois. Finalmente, os rastos e medidas existentes são classificados em três tipos: Rasto sem correspondência, medida sem correspondência, medida e rasto correspondentes.

O método acima descrito associa uma medida a um rasto não sendo capaz de lidar com junções e divisões, situações em que uma medida está associada a mais de um rasto e um rasto está associado a mais de uma medida. Por forma a resolver este problema é utilizado um processo de detecção de junções e divisões com base nos resultados obtidos.

3.3.2 Detecção de Junções e Divisões

Nos sistemas de vídeo vigilância é necessário ter métodos que permitam realizar associações um para vários e vários para um, uma vez que as regiões activas detectadas podem juntar-se ou dividir-se ao longo de uma sequência. Analisando a Figura 3.3 podemos verificar que da imagem #25 para a #40 o algoritmo de segmentação de fundo passa de detectar duas regiões activas para detectar apenas uma, é então necessário avaliar se as duas regiões activas em #25 se juntaram em apenas uma na imagem #40, ou se uma delas desapareceu. Da imagem #40 para a #80 acontece o contrário, passa-se de se detectar uma região activa em #40 para detectar duas em #80, é então necessário determinar se existe uma região nova em #80, ou a região activa da imagem #40 se dividiu em duas.

A secção anterior deixa rastos sem correspondência e medidas sem correspondência. Para os rastos sem correspondência, um algoritmo de detecção de junções é utilizado para decidir se o rasto se juntou a outra medida ou desapareceu. Se aconteceu uma junção,

é criado um grupo novo que contém a identidade dos indivíduos que estão sobrepostos. Se o rasto desapareceu, a confiança do rasto diminui, assim que descer abaixo de um determinado limiar, o rasto é apagado. Para as medidas sem correspondência é utilizado um módulo de detecção de divisões para decidir se a medida se separou de um rasto ou é um objecto novo.

Uma junção pode ocorrer devido a um rasto sem correspondência estar sobreposto a uma medida. Este julgamento baseia-se na ideia que tem que existir uma área sobreposta entre a *bounding box* que contém rasto e a *bounding box* que contém a medida, (ver Figura 3.4). Esta suposição é válida desde que o processo de segmentação seja suficientemente rápido, assim que os objectos se tocam em t, uma *bounding box* que contém os objectos que se juntaram é criada e tem grandes áreas sobrepostas às *bounding boxes* dos objectos em $t-1$. O método foi testado com sinais de vídeo adquiridos a um ritmo de $15fps$ e a hipótese manteve-se válida. O método de detecção de divisões é semelhante ao método de detecção de junções. Uma divisão é detectada devido a uma medida sem correspondência se sobrepor a um rasto, que se sabe conter duas pessoas sobrepostas Figura 3.4.

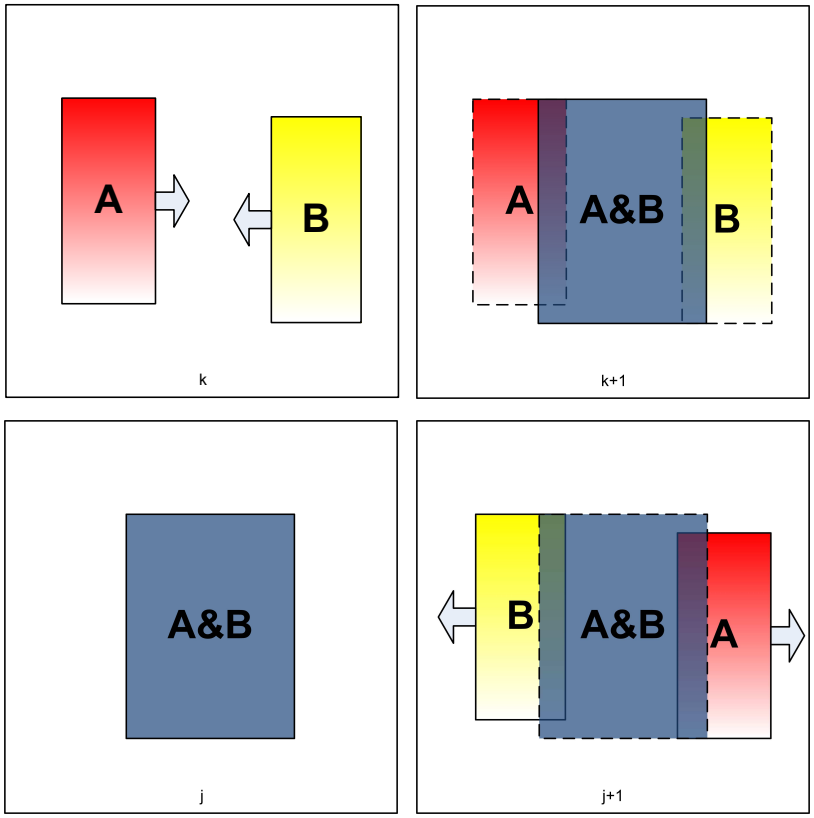


Figura 3.4: Exemplo de um cenário de detecção de uma junção e de uma divisão. A primeira linha mostra um cenário de junção, enquanto a segunda mostra uma divisão.

3.3.3 Resolução de Oclusões

Quando é detectada uma divisão de um rasto que se sabe conter duas pessoas sobrepostas, é necessário estabelecer a identidade das medidas resultantes da divisão. Para a realização desta tarefa é preciso guardar informação sobre cada pessoa seguida pelo sistema. Como muitas vezes as pessoas utilizam peças de roupa de cor diferente no tronco e nas pernas,

sempre que uma região activa j, contendo um indivíduo isolado, é detectada na imagem t, é extraído um descritor de características de cor c_j^t,

$$c_j^t = [\mu_{H_{sup}}, \mu_{S_{sup}}, \mu_{V_{sup}}, \mu_{H_{inf}}, \mu_{S_{inf}}, \mu_{V_{inf}}] \tag{3.9}$$

onde $\mu_{H_{sup}}$, $\mu_{S_{sup}}$, $\mu_{V_{sup}}$, $\mu_{H_{inf}}$, $\mu_{S_{inf}}$ e $\mu_{V_{inf}}$ correspondem às intensidades médias das componentes de cor H, S e V de todos os pixeis da metade superior e inferior, respectivamente, da *bounding box* da região activa j. Este descritor de características de cor c_j^t é de seguida utilizado para actualizar um descritor de características de cor C_i^t de cada indivíduo i detectado,

$$C_i^t = [\mu_{H_{Tsup}}, \mu_{S_{Tsup}}, \mu_{V_{Tsup}}, \mu_{H_{Tinf}}, \mu_{S_{Tinf}}, \mu_{V_{Tinf}}] \tag{3.10}$$

onde $\mu_{H_{Tsup}}$, $\mu_{S_{Tsup}}$, $\mu_{V_{Tsup}}$, $\mu_{H_{Tinf}}$, $\mu_{S_{Tinf}}$ e $\mu_{V_{Tinf}}$ correspondem às intensidades médias das componentes de cor H, S e V de todos os pixeis da metade superior e inferior da *bounding box* da região activa associada ao indivíduo i, para todas as imagens em que o indivíduo i se encontra isolado, desde a primeira imagem em que é detectado, até à imagem t.

Quando é detectada uma divisão na imagem t, é construída uma matriz de distâncias $\Delta^{C,c}$ entre os indivíduos i que estavam sobrepostos em $t-1$ e as regiões activas j presentes em t resultantes da divisão. A matriz de distâncias é baseada na distância Euclideana,

$$\Delta_{ij}^{C,c} = \sqrt{(C_i^t - c_j^t)(C_i^t - c_j^t)^\top}. \tag{3.11}$$

A associação entre os individuos i e as regiões activas j presentes em t pode então ser formulada como um problema de optimização e resolvido aplicando o algoritmo Húngaro [73].

3.4 Modelo Anatómico

No método em que se utiliza o modelo anatómico pretende-se, representar cada pessoa em cena por um modelo hierárquico do corpo humano. O modelo do corpo humano utilizado é semelhante ao proposto por [71] e consiste na divisão de cada indivíduo em cabeça, tronco e pernas e na subdivisão destas partes em zonas de pele e zonas sem pele, a Figura 3.5 ilustra a representação do corpo humano.

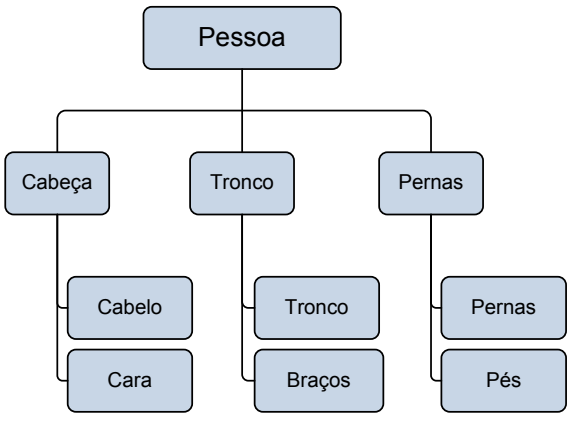


Figura 3.5: Modelo hierárquico do corpo humano.

Para construir o modelo hierárquico do corpo, cada região activa detectada é segmentada em *blobs*, agrupando pixeis indíviduais com base em intensidades de cor. Depois dos *blobs* serem atribuídos a cada indivíduo, os *blobs* são classificados em cabeça, tronco e pernas.

A construção do modelo do corpo é efectuada em todas as imagens das sequências em análise com vista à extracção de características descritivas das interacções que se pretende identificar. Paralelamente é mantido um modelo de cada indivíduo, que consiste na média da intensidade dos canais de cor H, S e V de cada parte anatómica para todas as imagens processadas até ao momento. Este modelo é utilizado para decidir a pertença de alguns *blobs* a cada indivíduo. É de seguida descrito cada módulo que foi desenvolvido no método do modelo anatómico.

3.4.1 Segmentação de Regiões Activas

O objectivo deste módulo consiste na segmentação de cada região activa detectada, em *blobs* constituídos por pixeis que partilhem características tanto de cor, como espaciais. Este módulo é constituído por 3 sub-módulos, detecção de pele, formação inicial de *blobs* e junção de *blobs* sobre-segmentados que se descrevem de seguida. A Figura 3.6 apresenta um diagrama de blocos do módulo de segmentação de regiões activas.

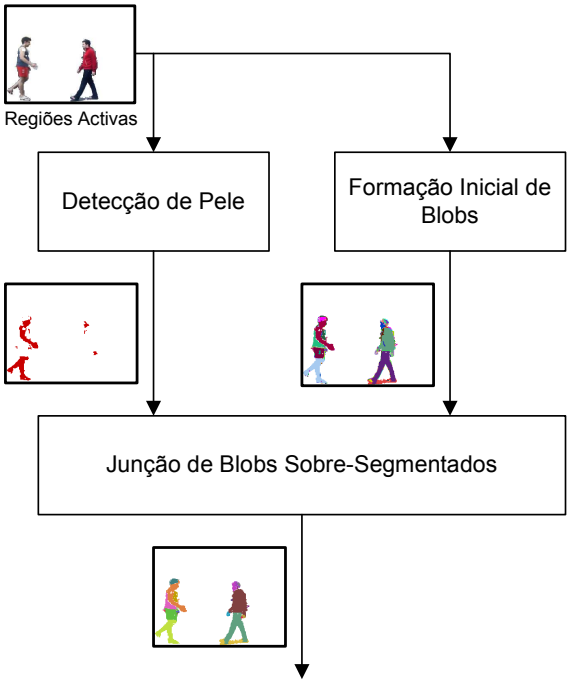


Figura 3.6: Diagrama de blocos da segmentação de regiões activas.

Detecção de Pele

A detecção de pele é extremamente útil no reconhecimento de partes do corpo. A cor da pele é determinada por um único pigmento (melanina), e apenas a sua densidade difere entre os diferentes grupos étnicos. Foi escolhido o espaço de cor RGB normalizado,

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B} \tag{3.12}$$

para representar a cor na detecção de pele, pois este espaço é invariante a mudanças de orientação das superfícies em relação à fonte de luz [74], e permite utilizar um modelo simples de limiares para detectar a cor da pele. Os limiares utilizados foram determinados por análise gráfica de conjuntos de pixeis de pele, como a Figura 3.7 ilustra, sendo depois sujeitos a testes para verificar a robustez dos mesmos. Assim, um pixel v pertencente a uma região activa, pertence também ao conjunto S dos pixeis de pele se se verificar a condição,

$$v \in S \text{ se } (\lambda_{g1}(r) \geq g \geq \lambda_{g2}(r)) \wedge (\lambda_{r1}(g) \geq r \geq \lambda_{r2}(g)) \tag{3.13}$$

com

$$\lambda_{g1}(r) = -0.3 \times r + 0.48 \tag{3.14}$$
$$\lambda_{g2}(r) = -042 \times r + 0.46 \tag{3.15}$$
$$\lambda_{r1}(g) = \frac{g - 0.237}{0.3} \tag{3.16}$$
$$\lambda_{r2}(g) = \frac{g + 0.13}{0.67} \tag{3.17}$$

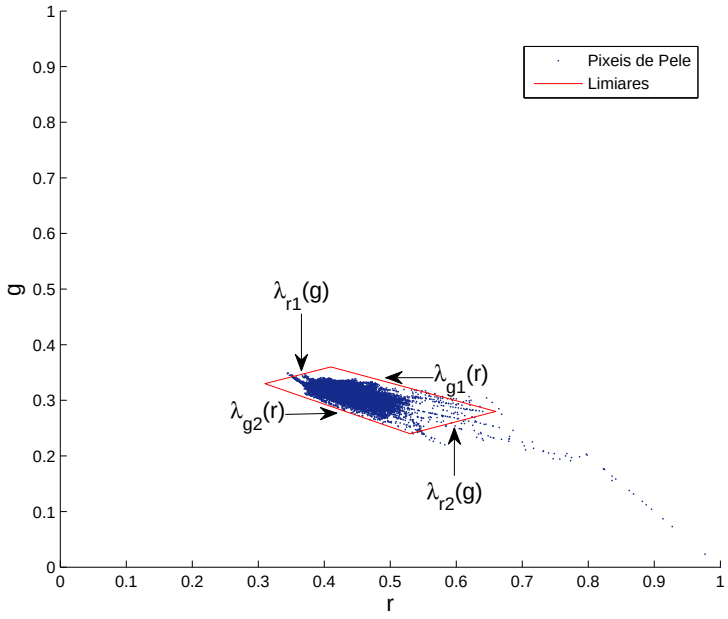


Figura 3.7: Exemplo de um conjunto de dados utilizado para determinar os limiares $\lambda_{g1}(r)$, $\lambda_{g2}(r)$, $\lambda_{r1}(g)$ e $\lambda_{r2}(g)$.

Formação Inicial de Blobs

O passo seguinte consiste em decompor cada região activa em sub-regiões com cor homogénea. Isso é feito admitindo que a cor de cada pixel $v(x, y)$ no espaço HSV é uma variável aleatória cuja distribuição é uma mistura de C Gaussianas. Assim, a cor de um pixel (x, y) no espaço HSV é uma variável aleatória $\mathbf{v} = [v_H, v_S, v_V]^\top$. Admitiremos que a distribuição probabilística dos pixels de foreground é uma mistura de C Gaussianas [75],

$$p(\mathbf{v}) = \sum_{r=1}^{C} p(\mathbf{v}|\omega_r)P(\omega_r), \tag{3.18}$$

em que

$$p(\mathbf{v}|\omega_r) = (2\pi)^{-d/2}|\Sigma_r|^{-1/2}e^{-\frac{(v-\mu_r)^T\Sigma_r^{-1}(v-\mu_r)}{2}}, \quad r = 1, ..., C. \tag{3.19}$$

e $P(\omega_r)$ é a probabilidade da r-ésima Gaussiana.

Cada componente θ_j da mistura é caracterizada pela probabilidade *a priori* $P(\omega_r)$, um vector das médias μ_r de cor dos pixeis, e uma matriz de covariâncias Σ_r. Assim, $\theta_j = \{P(\omega_j),\ \mu_j,\ \Sigma_j\},\ j = 1, ..., C$. Para se estimar os parâmetros da Gaussiana, é utilizado o algoritmo EM (esperança-maximização) [75]. Este algoritmo estima recursivamente os parâmetros da mistura a partir de estimativas iniciais, aplicando a seguinte actualização dos parâmetros

$$\hat{P}(\omega_i) \leftarrow \frac{1}{n}\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta}), \tag{3.20}$$

$$\hat{\mu}_i \leftarrow \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta})\mathbf{v}_k}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta})}, \tag{3.21}$$

$$\sum_{i=1}^{C} \leftarrow \frac{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta})(\mathbf{v}_k - \hat{\mu}_i)(\mathbf{v}_k - \hat{\mu}_i)^T}{\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta})} \tag{3.22}$$

onde

$$\hat{P}(\omega_i|\mathbf{v}_k, \hat{\theta}) \leftarrow \frac{p(\mathbf{v}_k|\omega_i, \hat{\theta}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{v}_k|\omega_j, \hat{\theta}_j)\hat{P}(\omega_j)} \tag{3.23}$$

O método é inicializado escolhendo a média aleatoriamente dentro da gama de valores presentes no *foreground* da imagem em análise para cada canal de cor H, S e V.

$$\mu_r = [v_H, v_S, v_V]^T$$

onde

$$\begin{aligned} v_H &\in [min(v_H), max(v_H)] \\ v_S &\in [min(v_S), max(v_S)] \\ v_V &\in [min(v_V), max(v_V)] \end{aligned} \tag{3.24}$$

A matriz de covariâncias é assumida como sendo uma matriz identidade.

$$\Sigma_r = diag(Var_1, Var_2, ..., Var_C) \tag{3.25}$$

Finalmente todas as probabilidades *a priori* são consideradas como sendo iguais.

$$P(\omega_r) = \frac{1}{C} \tag{3.26}$$

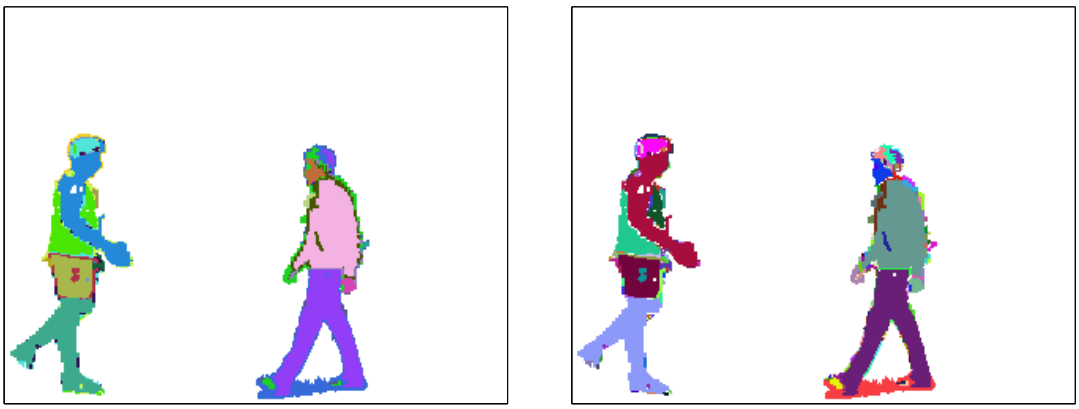
O treino é realizado actualizando iterativamente os parâmetros acima mencionados de acordo com as equações (3.20-3.22). O processo iterativo pára quando o valor da média muda menos de 1% quando comparado com o valor da iteração anterior, ou quando o

máximo de iterações permitidas definidas pelo utilizador ζ, é ultrapassado. Nesta tese usou-se ζ=200 e uma mistura de 10 Gaussianas ($C = 10$), cujos parâmetros são de seguida utilizados para classificar os pixeis numa das classes C utilizando um classificador de máximo à posteriori (MAP).

$$\omega_L = arg\ max_r\ log(P(\omega_r|\mathbf{v})), \quad 1 \leq r \leq C. \tag{3.27}$$

A probabilidade MAP $P(\omega_r|\mathbf{v})$ é calculada para todos os pixeis do *foreground* $\mathbf{v}$ e todas as classes r. A um pixel v é atribuída a marca de classe ω_L que produzir a maior probabilidade MAP.

O processo acima mencionado, classifica os pixeis com a mesma cor como sendo da mesma classe, mesmo que não estejam conectados. Os *blobs* que pretendemos obter são regiões conexas. Assim os *blobs* obtêm-se aplicando uma análise de regiões conexas aos pixeis classificados numa mesma classe, Figura 3.8. Este processo vai gerar um grande número de *blobs* que varia de acordo com a imagem de entrada. Torna-se então necessário proceder a uma análise que permita juntar *blobs* que se encontrem sobre-segmentados em regiões coerentes. Isto requer uma análise de alto-nível da imagem que tenha em conta as relações entre as regiões segmentadas.



(a) (b)

Figura 3.8: Comparação entre classificação inicial dos pixeis (a) e da formação inicial de *blobs* (b). O número de conjuntos passa de 16 em (a) para 128 em (b).

Junção de Blobs Sobre-Segmentados

A junção de blobs sobre-segmentados é um processo de crescimento de regiões [76] controlado por características dos própios blobs. Foram então extraídas dois tipos de características, primárias que dizem respeito ao próprio blob e secundárias que dizem respeito à relação entre blobs adjacentes, que permitem descrever o blob A_j como se segue:

1. Características primárias: referentes a uma única região
   - Marca do blob: $M(A_j) \in Z^+$.
   - Tamanho do blob: $\alpha(A_j) = |A_j|$, onde $|A_j|$ corresponde ao número de pixeis que pertencem ao blob.
   - Cor: $[\mu_H, \mu_S, \mu_V]^T$ as intensidades médias das componentes de cor H, S e V do blob.

- Posição do blob: $[\bar{I}, \bar{J}]^T$, isto é, a mediana das projecções horizontais e verticais do blob.
- Contorno do blob: $\Psi(A_j)$, conjunto dos pixeis que compõe a fronteira do blob.

2. Características secundárias: referentes a blobs adjacentes
   - Lista de adjacências: $\Gamma(A_j) = \{k \in Z^+ | A_k \text{ adjacente a } A_j\}\ k \neq j$.
   - Rácio da fronteira de A_j em relação a A_k: $\beta_j(A_k)$ =número de pixeis em $\Psi(A_j)$ conectados a A_k / $\Psi(A_j)$.
3. É incluído ainda o seguinte marcador de pele:
   - Marcador de pele: $\varsigma(A_j)$

$$\varsigma(A_j) = \begin{cases} 1 & \text{se } |A_j \cap S| \geq |A_j| \times 0.75 \\ 0 & \text{caso contrário} \end{cases}$$

   onde S é o conjunto de pixeis do foreground identificados como pixeis de pele.

Dois blobs A_j e A_r são então juntos se os seguintes critérios se verificarem:

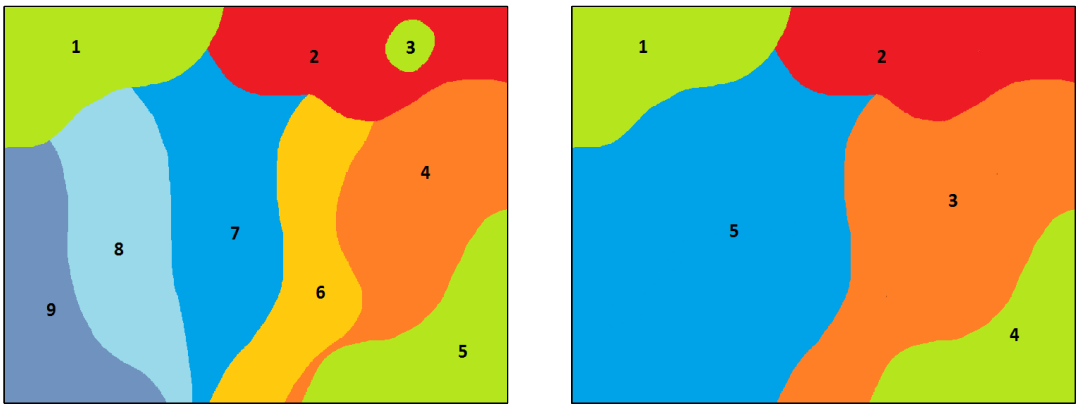
1. Critério de adjacência: Dois blobs só se juntam se forem adjacentes.
2. Critério da semelhança de cor: Dois blobs devem ter cor semelhante, onde a semelhança de cor é determinada pela distância de Mahalanobis δ_ϕ da característica de cor ϕ entre os blobs A_j e A_r como se segue:

$$\delta_\phi = (\phi_j - \phi_r)^T (\Sigma_\phi)^{-1} (\phi_j - \phi_r), \tag{3.28}$$

$$\phi = [\mu_H, \mu_S, \mu_V]^T, \tag{3.29}$$

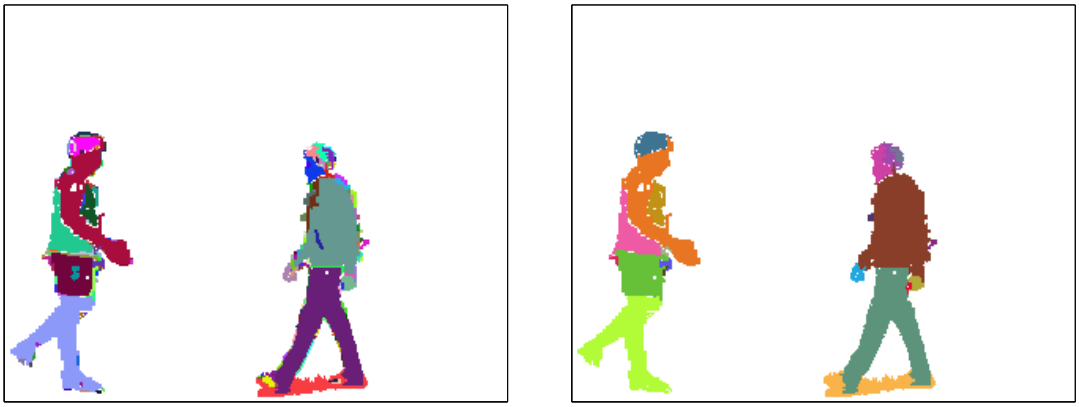
   onde Σ_ϕ é a matriz de covariâncias dos valores de cor de todas os blobs na imagem. Se δ_ϕ for menor que um determinado limiar λ_ϕ a cor de A_j e A_r é considerada semelhante.
3. Critério do rácio de fronteira: Dois blobs devem partilhar uma fronteira grande. $(\beta_j(A_r) \geq Th_\beta) \vee (\beta_r(A_j) \geq Th_\beta)$ onde Th_β é um limiar.
4. Critério do blob pequeno: Um blob A_j que seja menor que um limiar Th_α e que partilhe com outro blob A_r mais de 80% da sua fronteira $\beta_j(A_r) > 0.8$ não precisa de seguir o Critério 2 excepto se for um blob marcado como sendo pele $\varsigma(A_j) = 1$.

A figura 3.9 ilustra o processo de junção baseado nos critérios acima mencionados. Os blobs 4 e 6 e os blobs 7, 8 e 9 seguem os três primeiros critérios e são juntos num só blob. Já o blob 3 obedece ao critério quatro e é por isso junta ao blob 2. A figura 3.10 compara o *input* e *output* do módulo de junção de blobs sobre-segmentados. A figura 3.11 compara o número de blobs do *input* e *output* do módulo de junção de blobs sobre-segmentados ao longo de uma sequência dos dados experimentais. É de salientar a flutuação no número de blobs ao longo da sequência.



(a) (b)

Figura 3.9: Exemplo de um processo de junção de *blobs* sobre-segmentados.



(a) (b)

Figura 3.10: Comparação entre a entrada (a) e saída (b) do módulo de junção de *blobs* sobre-segmentados. O número de *blobs* passa de 128 em (a) para 21 em (b).

3.4.2 Gestão de Sobreposição

Quando é detectada uma junção no módulo de seguimento de regiões activas, este módulo é responsável pela segmentação de cada indivíduo sobreposto, enquanto durar a junção. A segmentação de cada indivíduo é realizada atribuindo a cada *blob*, pertencente à região activa que contém as pessoas sobrepostas, a identidade de uma delas. Para essa função foi criado um sistema composto por três módulos, correspondência com imagem anterior, classificação espacial e correspondência com modelo do corpo humano. Os *blobs* de cada silhueta em que se encontrem sobrepostos dois indivíduos, passam por cada um destes módulos até lhes ser atribuída a identidade de um dos indivíduos que se encontra sobreposto. Se no final não se conseguir atribuir a um *blob* a identidade de nenhum indivíduo ele é considerado não classificado e não é processado nas fases seguintes do sistema global. A Figura 3.12 ilustra a entrada e saída deste módulo. Descreve-se de seguida cada um dos sub-módulos que compõem o módulo de gestão de sobreposição.

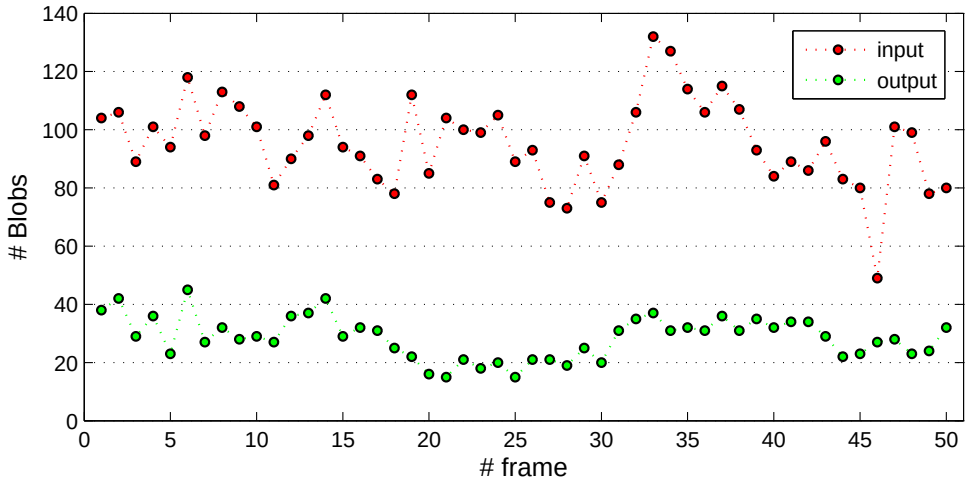
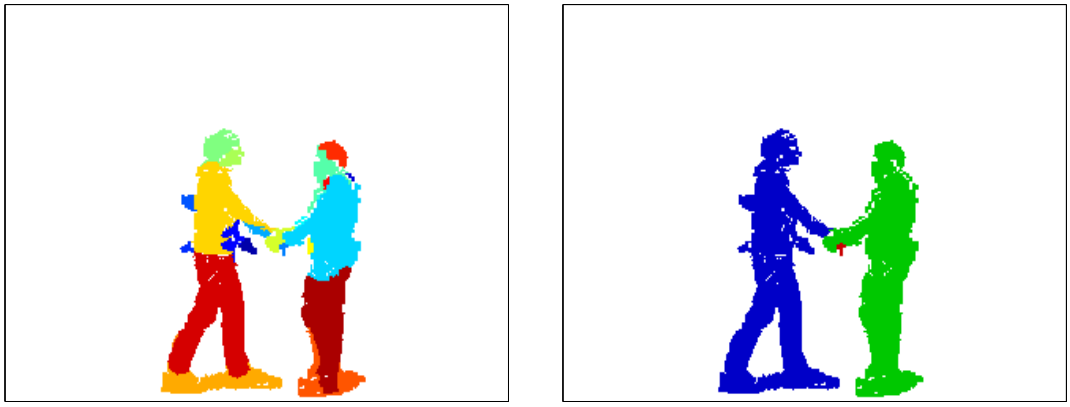


Figura 3.11: Comparação do número de *blobs* presentes na entrada e saída do módulo de junção de *blobs* sobre-segmentados.



(a) (b)

Figura 3.12: Exemplo da entrada do módulo de gestão de sobreposição (a) e a sua saída (b). Os *blobs* a vermelho não foram classificados.

Seguimento de Blobs

Existem *blobs* correspondentes a partes do corpo grandes, como tronco e pernas, cuja segmentação é robusta, mesmo quando dois indivíduos se encontram sobrepostos. Estes *blobs* são então fáceis de seguir durante o tempo em que as pessoas se encontram sobrepostas. Assim o primeiro passo na identificação da identidade dos *blobs*, consiste num processo de seguimento dos mesmos. Este processo de seguimento envolve alguns problemas:

1. Existência de um diferente número de *blobs* em cada imagem.
2. Um único *blob* em $t-1$ pode dividir-se em vários em t, devido a sombras, oclusões, etc.
3. Vários *blobs* em $t-1$ podem juntar-se em t, devido a sobreposições, oclusões, etc.
4. Alguns *blobs* em $t-1$ podem desaparecer em t.

5. Novos *blobs* podem aparecer em t.

6. É necessário manter a identidade dos *blobs* ao longo da sequência.

Estes fenómenos complicam o seguimento de *blobs*, é preciso permitir associações de vários blobs em $t-1$ a um único *blob* em t e vice versa. É preciso no entanto evitar situações em que este tipo de associação é indesejado, como por exemplo um *blob* em $t-1$ ser associado a vários *blobs* em t que se encontram dispersos.

Para a realização desta tarefa foi utilizada uma adaptação de um algoritmo de seguimento múltiplo proposto em [77]. Utilizou-se um método de seguimento distinto do já apresentado para o seguimento de regiões activas, pois no caso do seguimento de *blobs* era necessário ter um método que fosse capaz de dar mais ênfase aos *blobs* maiores e que lidasse com as junções e separações de forma diferente. Vamos chamar rastos R^{t-1} ao conjunto de *blobs* presentes em $t-1$ e B^t ao conjunto de novos *blobs* formados em t. O rasto número i é $R_i^{t-1} \in R^{t-1}$ e o *blob* j é $B_j^t \in B^t$.

O seguimento de *blobs* é associar um *blob* B_j^t na imagem t a um ou mais rastos R^{t-1} presentes na imagem $t-1$. A correspondência entre os dois conjuntos R^{t-1} e B^t é representada pela matriz binária E,

$$E^{t-1,t} = \begin{pmatrix} \epsilon_{11}^{t-1,t} & \epsilon_{12}^{t-1,t} & \dots \\ \epsilon_{21}^{t-1,t} & \epsilon_{22}^{t-1,t} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.30}$$

onde

$$\epsilon_{ij}^{t-1,t} = \begin{cases} 1 \text{ se o rasto } R_i^{t-1} \text{ é associado ao } \textit{blob } B_j^t \\ 0 \text{ caso contrário} \end{cases} \tag{3.31}$$

A associação entre *blobs* e rastos é realizada comparando a semelhança entre vectores de características descritivos de cada *blob* m_i^{t-1} e m_j^t,

$$m_i^{t-1} = [\alpha, \mu_H, \mu_S, \mu_V, \bar{I}, \bar{J}] \text{ para } R_i^{t-1}, \tag{3.32}$$

$$m_j^t = [\alpha, \mu_H, \mu_S, \mu_V, \bar{I}, \bar{J}] \text{ para } B_j^t, \tag{3.33}$$

onde α corresponde à área do *blob*, μ_H, μ_S e μ_V são as intensidades médias das componentes de cor H, S e V do *blob* e $\bar{I}$, $\bar{J}$ são as medianas das projecções horizontais e verticais do *blob*.

Dadas as matrizes de covariâncias Π_{t-1} e Π_t destas características para todos rastos presentes na imagem em $t-1$ e todos os *blobs* presentes em t, respectivamente, a distância de Mahalanobis $\Delta_{ij}^{t-1,t}$ entre o rasto número i, R_i^{t-1} presente em $t-1$ e o *blob* número j, B_j^t presente em t é dada por:

$$\Delta_{ij}^{t-1,t} = (m_i^{t-1} - m_j^t)^T (\Pi_{t-1} + \Pi_t)^{-1} (m_i^{t-1} - m_j^t). \tag{3.34}$$

O processo de seguimento é composto por duas fases, uma primeira fase onde só se permite realizarem associações um para um, e uma segunda onde são realizadas as associações um para vários.

A primeira fase tem como objectivo associar os rastos e os *blobs* maiores, correspondentes a partes do corpo grandes e homogéneas como tronco, pernas, braços, etc. de forma a tornar o sistema robusto contra oclusões parciais, sombras e ruído. Esta associação inicial está sujeita às restrições de associação um para um:

$$\sum_{j=1}^{|B^t|} \epsilon_{ij}^{t-1,t} = 1,\ \forall i = 1, ..., |R^{t-1}| \tag{3.35}$$

$$\sum_{i=1}^{|R^{t-1}|} \epsilon_{ij}^{t-1,t} = 1,\ \forall j = 1, ..., |B^t| \tag{3.36}$$

Esta primeira fase do processo de correspondências vai criar quatro subconjuntos, ${}^1R^{t-1} \subseteq R^{t-1}$, ${}^0R^{t-1} \subseteq R^{t-1}$, ${}^1B^t \subseteq B^t$ e ${}^0B^t \subseteq B^t$ correspondentes aos conjuntos de rastos e *blobs* com e sem correspondência respectivamente. Descreve-se de seguida o processo utilizado para realizar a associação um para um:

1. Calculo da matriz de diferenças $\Delta^{t-1,t}$:

$$\Delta^{t-1,t} = \begin{pmatrix} \Delta_{11}^{t-1,t} & \Delta_{12}^{t-1,t} & ... \\ \Delta_{21}^{t-1,t} & \Delta_{22}^{t-1,t} & ... \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.37}$$

   Se o valor de $\Delta_{ij}^{t-1,t}$ for maior que um determinado limiar Th_Δ o seu valor é actualizado para infinito.

2. Procura para trás com auxilio de $\epsilon_{Bij}^{t-1,t}$:
   Procurar para cada *blob* $j \in B^t$, por ordem descendente de tamanho o rasto, $i \in \{1, ..., |R^{t-1}|\}$ que tenha a diferença $\Delta_{ij}^{t-1,t} < Th_\Delta$ mínima, respeitando a restrição da equação 3.35.

$$\epsilon_{Bij}^{t-1,t} = \begin{cases} 1 \text{ se existir algum rasto } i \text{ que respeite as condições.} \\ 0 \text{ caso contrário.} \end{cases}$$

3. Procura para a frente com auxilio de $\epsilon_{Rij}^{t-1,t}$:
   Procurar para cada rasto $i \in R^{t-1}$, por ordem descendente de tamanho o blob, $j \in \{1, ..., |B^t|\}$ que tenha a diferença $\Delta_{ij}^{t-1,t} < Th_\Delta$ mínima, respeitando a restrição da equação 3.36.

$$\epsilon_{Rij}^{t-1,t} = \begin{cases} 1 \text{ se existir algum } blob\ j \text{ que respeite as condições.} \\ 0 \text{ caso contrário.} \end{cases}$$

4. Combinar procuras para a frente e para trás:

$$\epsilon_{ij}^{t-1,t} = \begin{cases} 1 \text{ se } (\epsilon_{Bij}^{t-1,t} = 1) \wedge (\epsilon_{Rij}^{t-1,t} = 1). \\ 0 \text{ caso contrário.} \end{cases}$$

A primeira fase do processo de seguimento de *blobs* deixa sem correspondência alguns rastos ${}^0R^{t-1}$ e alguns *blobs* ${}^0B^t$. Isto deve-se ao facto de esta primeira fase só permitir que sejam realizadas associações um para um. Por exemplo, assumindo que dois *blobs* têm a sua diferença mínima com um determinado rasto, apenas o *blob* que tenha a diferença mais baixa é associada ao rasto. O segundo *blob* fica retido para a próxima fase da associação como um *blob* residual. Realizam-se as seguintes associações depois da primeira fase:

1. Associação entre ${}^1R^{t-1}$ e ${}^0B^t$.

2. Associação entre ${}^1B^{t-1}$ e ${}^0R^t$.

Não são feitas associações entre ${}^0R^{t-1}$ e ${}^0B^t$, pois se existisse alguma associação, esta já teria sido estabelecida na associação um para um prévia, como uma associação entre ${}^1R^{t-1}$ e ${}^1B^t$.

O processo utilizado para realizar as associações nesta fase consiste em procurar para cada rasto $R_i^{t-1} \in {}^0R^{t-1}$ o *blob* $B_j^t \in {}^1B^t$ que tenha a diferença $\Delta_{ij}^{t-1,t}$ mínima, e procurar para cada *blob* $B_j^t \in {}^0B^t$ o rasto $R_i^{t-1} \in {}^1R^{t-1}$ que tenha também a diferença $\Delta_{ij}^{t-1,t}$ mínima e realizar a respectiva associação. No entanto nem todos os elementos de ${}^1R^{t-1}$ e ${}^1B^t$ são passíveis de serem associados a elementos de ${}^0B^t$ e ${}^0R^{t-1}$ respectivamente. Um rasto $R_{i=a}^{t-1} \in {}^0R^{t-1}$ só pode ser associado a um *blob* $B_{j=b}^t \in {}^1B^t$ se o rasto $R_{i=c}^{t-1} \in {}^1R^{t-1}$ a que o *blob* já está associado, $\epsilon_{i=c,j=b}^{t-1,t} = 1$, estiver contido na sua lista de adjacências $R_{i=c}^{t-1} \subseteq \Gamma(R_{i=a}^{t-1})$. No final deste processo podem existir quatro resultados possíveis para cada *blob* $B_j^t \in B^t$,

- Um *blob* é associado a um rasto.
- Um *blob* é associado a vários rastos.
- Vários *blobs* são associados a um rasto.
- Um *blob* não é correspondido.

Para cada uma destas situações é necessário avaliar se é possível atribuir a cada *blob* a identidade de um dos indivíduos com base no processo de seguimento. Caso contrário, o *blob* é marcado para processamento nos passos seguintes do módulo de gestão de sobreposição. Assim, para um *blob* que tem uma correspondência directa com um rasto, só é possível atribuir-lhe uma identidade caso a identidade do rasto a que está associado for conhecida. No caso em que um *blob* está associado a mais de um rasto, apenas se a identidade dos vários rastos for conhecida e esta for igual para todos os rastos é possível atribuir-lhe uma identidade. Quando um rasto está associado a vários *blobs*, apenas se a identidade do rasto for conhecida, todos os *blobs* a que este está associado podem herdar a sua identidade. No caso em que um *blob* não é correspondido é impossível atribuir-lhe uma identidade, sendo por isso processado nos passos seguintes. Podemos então considerar que o processo de seguimento de *blobs* divide o conjunto B^t em três sub-conjuntos I^k, com $k = 0, 1, 2$. Onde I^0 corresponde ao conjunto de *blobs* a que não foi atribuída identidade, e I^1 e I^2 correspondem aos conjuntos de *blobs* a que foram atribuídas as identidades de cada um dos sujeitos.

Classificação Espacial

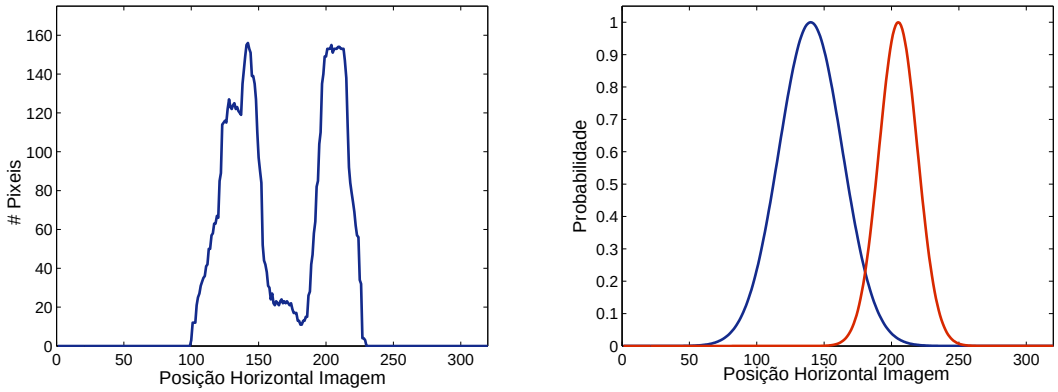
O módulo de seguimento de *blobs* nem sempre é capaz de atribuir a identidade de uma das pessoas sobrepostas a todos os *blobs*. Com o fim de classificar os *blobs* $B_j \in I^0$ é realizado um processo de classificação espacial, que se baseia no facto de em grande parte das situações em que uma região activa contém duas pessoas sobrepostas, estas mantêm alguma distância entre elas e têm apenas pequenas partes do corpo em contacto.

Para realizar a classificação espacial dos *blobs*, primeiro é realizada a projecção vertical da região activa que contém as duas pessoas sobrepostas para de seguida se treinar uma mistura de duas Gaussianas uni-dimensionais utilizando o algoritmo EM. A Figura

3.13(a) mostra um exemplo da projecção da região activa da Figura 3.12(a), a Figura 3.13 mostra a mistura de Gaussianas treinada. O passo seguinte, consiste em atribuir a cada uma das Gaussianas estimadas a identidade de uma das pessoas sobrepostas. Nesse sentido é calculada para cada Gaussiana $\mathcal{N}_i$, a probabilidade de pertencer ao indivíduo k $P(\mathcal{N}(\mu_i, \sigma_i) \Rightarrow k)$ dada por,

$$P(\mathcal{N}(\mu_i, \sigma_i) \Rightarrow k) = \frac{1}{N_{I^k}} \sum_{l=1}^{N_{I^k}} \frac{1}{N_{B_l}} \sum_{r=1}^{N_{B_l}} x_r \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(x_r-\mu_i)^2}{2\sigma_i^2}} \quad com \;\; i = 1,2 \;\; e \;\; k = 1,2 \tag{3.38}$$

onde N_{I^k} corresponde ao número de *blobs* pertencentes a I^k, N_{B_l} corresponde ao número de pixeis do *blob* $B_l \in I^k$ e x_r à posição horizontal do r-ésimo pixel pertencente ao *blob* l. Depois de calculada cada uma destas probabilidades a associação entre cada Gaussiana i e cada indivíduo k pode então ser formulada como um problema de optimização e resolvido utilizando o algoritmo Húngaro [73].



(a) (b)

Figura 3.13: Projecção da região activa da Figura 3.12 (a) e a mistura de Gaussianas treinada (b).

Uma vez resolvido o problema de associação entre as Gaussianas e os indivíduos pode-se proceder à atribuição da identidade de um dos sujeitos a cada *blob* $B_j \in I^0$. Para isso é calculada a probabilidade de cada *blob* B_j pertencer ao indivíduo k dada por,

$$P(B_j \Rightarrow k) = \frac{1}{N_{B_j}} \sum_{m=1}^{N_{B_j}} x_m \frac{1}{2\pi\sigma_k} e^{-\frac{(x_m-\mu_k)^2}{2\sigma_k^2}} \quad com \;\; k = 1,2 \tag{3.39}$$

onde N_{B_j} corresponde ao número de pixeis do *blob* B_j e x_m à posição horizontal do m-ésimo pixel de B_j. Uma vez calculada a probabilidade de um *blob* pertencer a cada indivíduo k, este é associado ao indivíduo que tiver a maior probabilidade. Desde que esta seja maior que 0.7 e que o valor absoluto da diferença entre as duas probabilidades seja superior a 0.3, ou seja um *blob* só é associado a um indivíduo caso a probabilidade de pertencer a este for elevada e a probabilidade de pertencer ao outro indivíduo for bastante inferior. Caso estas condições não se verifiquem o *blob* é dado como não classificado e marcado para processamento no passo seguinte do módulo de gestão de sobreposição. Na prática a classificação espacial apenas é capaz de efectivamente atribuir a identidade de um indíviduo aos *blobs*, quando as pessoas sobrepostas mantêm alguma distância entre

elas, uma vez que quando um indíviduo oclude completamente o outro, ou quando as pessoas têm grandes áreas do corpo em contacto, a probabilidade de um *blob* pertencer a um indivíduo é semelhante para ambas as pessoas.

Correspondência com Modelo do Corpo

Os *blobs* que não foram classificados nos passos prévios do módulo de gestão de sobreposição, passam por um último passo de processamento na tentativa de os classificar. Com o fim de identificar estes *blobs*, é realizada uma comparação entre um vector de características b_j de cada *blob* j que ainda não foi classificado e cada vector de características v_i^k que compõe o modelo do corpo de cada indivíduo k,

$$b_j = [\mu_H(b),\ \mu_S(b),\ \mu_V(b)] \tag{3.40}$$

$$v_i^k = [\mu_H(v),\ \mu_S(v),\ \mu_V(v)] \tag{3.41}$$

onde μ_H, μ_S e μ_V correspondem às intensidades médias de cada componente de cor H, S e V.

Dadas as matrizes de covariâncias Π_b e Π_v destas características para todos os blobs sem correspôndencia e todas as caracaterísticas que compõe o modelo de cada indivíduo, respectivamente, a distância de Mahalanobis Δ_{ij} entre o blob número j e cada característica i é dada por,

$$\Delta_{ij} = (b_j v_i^k)^T (\Pi_b + \Pi_v)^{-1} (b_j v_i^k). \tag{3.42}$$

Cada valor de Δ_{ij} é comparado com um limiar, caso o seu valor seja inferior é realizada a respectiva associação e atribuição de identidade. Neste processo são excluídos os vectores de características do modelo de cada indíviduo que descrevem zonas de pele, uma vez que estas zonas não permitem fazer uma distinção entre indíviduos.

Os blobs que não são identificados neste módulo são considerados não classificados e não são processados nos módulos seguintes do sistema.

3.4.3 Atribuição Partes Anatómicas

Neste módulo pretende-se segmentar a silhueta de cada indivíduo em três partes, cabeça, tronco e pernas. Pretende-se ainda sub-dividir cada uma destas três partes em zonas de pele e zonas sem pele. Isto é conseguido classificando cada um dos *blobs* que pertencem a determinado indivíduo numa das três partes que se pretende segmentar. O processo de classificação dos *blobs* começa dividindo a *bounding box* que contém a silhueta de cada pessoa em três zonas: cabeça, tronco e pernas. Esta divisão é feita com base no conhecimento de proporções básicas da anatomia humana, assim:

- A zona da cabeça é definida como o intervalo vertical entre o topo da *bounding box* e κ_1 vezes a altura h da silhueta da pessoa.
- A zona do tronco é definida como o intervalo vertical entre κ_1 vezes a altura e κ_2 vezes a altura da silhueta.
- A zona das pernas é definida como o restante intervalo vertical da silhueta.

A Figura 3.14 mostra um exemplo de divisão da *bounding box*. Após se ter realizado a divisão, é calculado para cada *blob* a percentagem da sua área que se encontra em cada uma das diferentes zonas da *bounding box*. Os *blobs* são de seguida classificados com base na zona da *bounding box* onde têem a maior parte da sua área. Ou seja, um *blob* que

tenha por exemplo 80% da sua área na zona da cabeça e os restantes 20% na zona do tronco é classificado como pertencendo à cabeça do indivíduo. A sub-divisão de cada uma das partes anatómicas é realizada com base num *blob* ter sido identificado como um *blob* de pele no módulo de segmentação de regiões activas. Assim, um *blob* que tenha sido classificado como pertencendo à cabeça de uma pessoa é classificado como cara caso tenha sido identificado como um *blob* de pele anteriormente, caso contrário é classificado como cabelo.

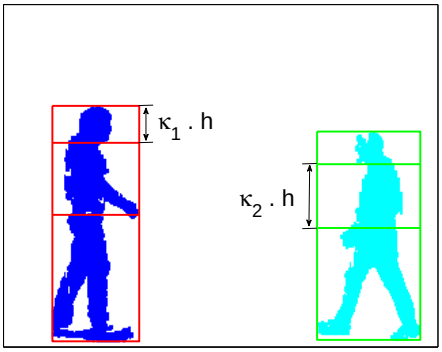


Figura 3.14: Exemplo da divisão da *bounding box* que contém a silhueta de cada pessoa.

Os valores de κ_1 e κ_2 são inicializados a 0.16 e 0.45 respectivamente. A escolha do valor de inicialização de cada um destas parâmetros foi feita com base no conhecimento das proporções do corpo humana e afinada com base nos resultados experimentais. Os parâmetros κ_1 e κ_2 são actualizados dinâmicamente de acordo com a mudança de aparência de cada pessoa em cada imagem, sendo depois os valores actualizados, utilizados como valores iniciais para estimar as partes do corpo na próxima imagem.

3.5 Resultados Experimentais

3.5.1 Subtracção de Fundo

A aproximação tradicional para avaliar um resultado da segmentação de fundo é compararálo com uma segmentação de referência. Nesse sentido seleccionou-se manualmente os pixels activos em 100 imagens do conjunto de dados, contendo tanto pessoas isoladas como sobrepostas. De seguida para fazer a avaliação do desempenho do algoritmo foram utilizados dois métodos. O primeiro método consiste em identificar e quantificar os erros de tipo I (falsos positivos) e tipo II (falsos negativos). O segundo método utilizado foi a métrica de Hammoude [78], definida como

$$H = \frac{(A \cup \tilde{A}) - (A \cap \tilde{A})}{(A \cup \tilde{A})} \tag{3.43}$$

onde, o conjunto A corresponde ao conjunto de pixeis activos seleccionados manualmente e o conjunto $\tilde{A}$ ao conjunto de pixeis activos obtidos pelo algoritmo. A métrica de Hammoude tem valores entre 0 e 1, com o valor 1 a ocorrer quando não existe intersecção entre os conjuntos e o valor 0 a ocorrer quando os conjuntos são iguais. A Tabela 3.1 apresenta os resultados para as duas métricas utilizadas. A Figura 3.15 ilustra quatro exemplos de imagens de entrada e os respectivos *foregrounds* reais e estimados, bem como a classificação dos pixeis.

Tabela 3.1: Resultados da Subtracção de Fundo

	μ (σ)
Erros tipo I (%)	0.76 (0.26)
Erros tipo II (%)	0.87 (0.33)
Métrica de Hammoude	0.15 (0.04)

Analisando os resultados, podemos concluir que o módulo de segmentação de fundo tem um bom desempenho. A métrica de Hammoude é baixa e os erros do tipo I e II são inferiores a 1%. Os erros do tipo I como pode ser verificado na Figura 3.15 devem-se sobretudo a borrões causados pelo movimento nas extremidades dos indivíduos e às sombras que os indivíduos projectam no chão. Existem algoritmos capazes de minimizar a detecção de sombras [15], no entanto face aos resultados apresentados não se achou necessário aplicar esse tipo de técnicas. Os erros do tipo II devem-se à cor do *foreground* ser muito parecida com a do fundo. Isto poderia ser resolvido utilizando uma camera com uma maior resolução ao nível da cor, ou incorporando outras características nas segmentação para além da cor, como por exemplo a textura, ou o movimento.

3.5.2 Detecção de Pele

Por forma a avaliar a detecção de pele recorreu-se aos mesmos métodos utilizados para a avaliação da segmentação de fundo. Nesse sentido foram seleccionados manualmente em 20 imagens pertencentes ao conjunto de dados, todos os pixeis de pele pertencentes ao *foreground* detectado em cada uma das imagens. A Tabela 3.2 apresenta os resultados para as duas métricas utilizadas. A Figura 3.16 ilustra quatro exemplos de entrada do módulo de detecção de pele e os respectivos pixeis de pele reais e estimados, bem como a classificação dos pixeis.

Tabela 3.2: Resultados da Detecção de Pele.

	μ (σ)
Erros tipo I (%)	6.57 (3.29)
Erros tipo II (%)	3.41 (1.47)
Métrica de Hammoude	0.11 (0.07)

Analisando os resultados podemos concluir que o desempenho do módulo de detecção de pele é razoável, indo ao encontro dos resultados apresentados pela maioria dos métodos de detecção de pele [74]. Os erros do tipo I e II são baixos, não atingindo a ordem de grandeza das dezenas e a métrica de Hammoude apresenta também um resultado baixo. Analisando a Figura 3.16 podemos verificar que os erros do tipo I se devem sobretudo à projecção de sombras em zonas de pele. Os erros do tipo II devem-se sobretudo a este método não ser capaz de descriminar com eficácia a cor amarela, como não sendo pele. Este problema poderia ter sido resolvido utilizando não só o espaço de cor RGB normalizado na detecção de pele, mas também o espaço de cor HSV. Optou-se por utilizar apenas o espaço de cor RGB normalizado, pois utilizando também o espaço de cor HSV os erros do tipo I iriam sofrer um aumento.

Figura 3.15: Exemplo avaliação segmentação de fundo: (1ª linha) imagem de entrada, (2ª linha) *foreground* real, (3ª linha) *foreground* calculado, (4ª linha) classificação dos pixeis, onde as cores verde, branco, vermelho e amarelo correspondem a pixeis verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente.

3.5.3 Gestão de Sobreposição

Para avaliar o desempenho deste módulo, foram utilizadas 60 imagens de dois indivíduos sobrepostos, pertencentes ao conjunto de dados, onde se seleccionou manualmente a segmentação real dos mesmos. De maneira a apresentar os resultados deste módulo, sem propagar os erros do módulo de segmentação de fundo, a segmentação real foi manualmente seleccionada sobre o *foreground* estimado.

O módulo de gestão de sobreposição estima dois conjuntos, $\tilde{A}_1$ e $\tilde{A}_2$, compostos pelos pixeis correspondentes a cada um dos indivíduos presentes em cena. No entanto, por vezes, o módulo não é capaz de atribuir a todos os pixeis uma classificação, gerando por isso um terceiro conjunto, composto por pixeis não classificados. Nesse sentido, e por forma a avaliar o desempenho do módulo, foram utilizadas as métricas *precision* e *recall*, onde a *precision* reflecte uma medida da fiabilidade dos resultados, e o *recall* pode ser visto como uma medida da sensibilidade. A *precision* é dada por,

$$Precision = \frac{\#((A_1 \cap \tilde{A}_1) \cup (A_2 \cap \tilde{A}_2))}{\#(\tilde{A}_1 \cup \tilde{A}_2)} \tag{3.44}$$

onde A_1 e A_2 correspondem ao conjunto de pixeis da segmentação real de cada indivíduo, e $\tilde{A}_1$ e $\tilde{A}_2$ correspondem ao conjunto de pixeis da segmentação estimada de cada indivíduo.

Figura 3.16: Exemplo avaliação da detecção de pele: (1ª linha) entrada, (2ª linha) pele real, (3ª linha) pele detectada, (4ª linha) classificação dos pixeis, onde as cores verde, azul, vermelho e amarelo correspondem a verdadeiro positivo, verdadeiro negativo, falso negativo e falso positivo, respectivamente.

O *Recall* é dado por,

$$Recall = \frac{\#((A_1 \cap \tilde{A}_1) \cup (A_2 \cap \tilde{A}_2))}{\#(A_1 \cup A_2)} \tag{3.45}$$

Uma vez que a entrada deste módulo varia conforme a inicialização do algoritmo EM no módulo de formação de *blobs*, que é feita de forma aleatória, foram realizados 5 testes para cada imagem da segmentação real, perfazendo um total de 300 testes.

A tabela 3.3 apresenta os resultados para as duas métricas utilizadas, bem como a percentagem de pixeis não classificados. A Figura 3.17 ilustra quatro exemplos de entrada do módulo de gestão de sobreposição e as respectivas segmentações reais e estimadas, bem como uma representação dos pixeis classificados, correctamente e incorrectamente, e não classificados.

A análise dos resultados da gestão de sobreposição, permite-nos concluir que este módulo tem um desempenho razoável. A *precision* é bastante elevada, e o facto de o *recall* ser da mesma ordem de grandeza da percentagem dos pixeis classificados (100%-%pixeis não classificados), indica que a maior parte dos erros se deve a não se conseguir classificar alguns *blobs*. Os *blobs* que são classificados, são correctamente classificados. Os *blobs* que não são classificados, são sobretudo *blobs*, que contêm partes dos dois indivíduos, o

Tabela 3.3: Resultados da Gestão de Sobreposição.

	μ (σ)
Precision (%)	88.9 (13.5)
Recall (%)	78.1 (21.1)
Pixeis não classificados (%)	13.1 (18.1)

que torna impossível a sua correcta classificação. Isto acontece quando dois indivíduos se encontram muito próximos um do outro, o que implica que no módulo de segmentação de regiões activas sejam criados *blobs* que pertencem a mais de um indivíduo. Uma conclusão que os resultados experimentais não reflectem, pois as amostras avaliadas foram escolhidas aleatoriamente, mas que a observação experimental permite tirar, é que quanto mais tempo dura a sobreposição, piores são os resultados da gestão de sobreposição.

3.5.4 Atribuição Partes Anatómicas

Para avaliar o desempenho deste módulo, e à semelhança do que foi feito nas secções anteriores, foram seleccionadas manualmente as segmentações reais em 100 imagens do conjunto de dados. Destas 100 imagens, 40 contêm duas pessoas isoladas e 60 contêm duas pessoas sobrepostas. Mais uma vez de maneira a apresentar os resultados deste módulo, sem propagar os erros do módulo de segmentação de fundo, a segmentação real foi manualmente seleccionada sobre o *foreground* estimado.

Foram utilizados dois métodos para avaliar a atribuição de partes anatómicas. O primeiro consiste na construção de uma matriz de confusão que mostre a classificação das diversas partes anatómicas. O segundo método utiliza uma métrica baseada na métrica de Hammoude e avalia a atribuição das partes anatómicas de um modo global. Esta métrica é dada por

$$H = 1 - \frac{\#((C \cap \tilde{C}) \cup (T \cap \tilde{T}) \cup (P \cap \tilde{P}))}{\#(C \cup T \cup P)} \tag{3.46}$$

onde os conjuntos C, T e P correspondem aos conjuntos de pixeis da segmentação real da cabeça, tronco e pernas respectivamente. Os conjuntos $\tilde{C}$, $\tilde{T}$ e $\tilde{P}$ correspondem aos conjunto de pixeis da segmentação estimada.

No caso das pessoas sobrepostas foram ainda realizados ensaios, onde a segmentação real foi seleccionada apenas sobre os pixeis correctamente classificados no módulo de gestão de sobreposição, afim de se ter uma avaliação que não propagasse os erros deste módulo.

Uma vez que a entrada do módulo de atribuição de partes anatómicas varia conforme a inicialização do algoritmo EM no módulo de formação de *blobs*, que é feita de forma aleatória, foram realizados 5 testes para cada imagem da segmentação real, perfazendo um total de 500 testes, 200 para pessoas isoladas e 300 para pessoas sobrepostas. Optou-se por apresentar os resultados de pessoas isoladas e sobrepostas separadamente, pois quando as pessoas presentes em cena se encontram sobrepostas, existe um passo a mais de processamento, o que afecta bastante o desempenho deste módulo.

As tabelas 3.4, 3.5 e 3.6 apresentam os resultados das partes anatómicas individuais para pessoas isoladas e pessoas sobrepostas com e sem erros provenientes do módulo de gestão de sobreposição. A tabela 3.7 apresenta os resultados da avaliação global da atribuição de partes anatómicas. As figuras 3.18 e 3.19 ilustram quatro exemplos da entrada do módulo de atribuição de partes anatómicas as segmentações reais e estimadas,

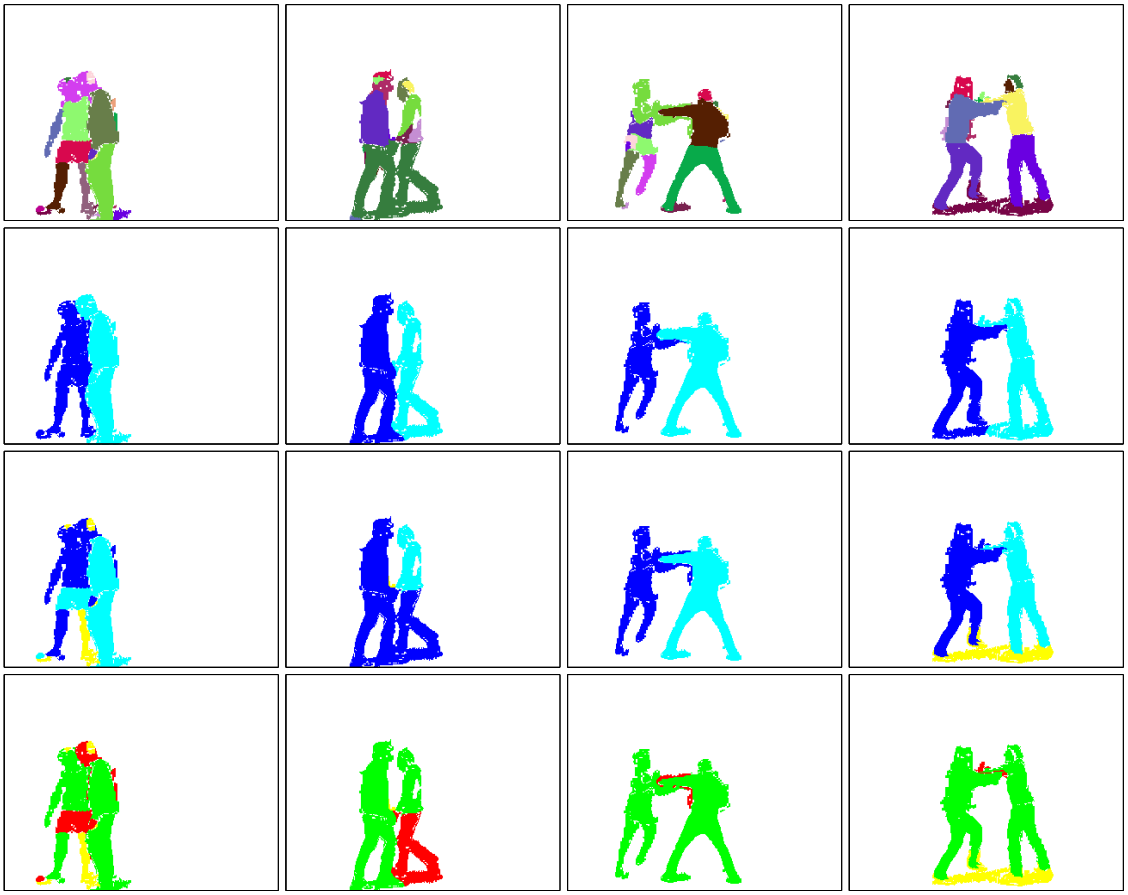


Figura 3.17: Exemplo avaliação da gestão de sobreposição: (1ª linha) entrada, (2ª linha) segmentação real, (3ª linha) segmentação estimada, (4ª linha) classificação dos pixeis, onde as cores verde, vermelho e amarelo correspondem a pixeis , correctamente classificados, mal classificados e não classificados, respectivamente.

bem como uma representação da classificação dos pixeis para pessoas isoladas e sobrepostas, respectivamente.

Tabela 3.4: Matriz de confusão da atribuição de partes anatómicas para pessoas isoladas.

	Cabeça	Tronco	Pernas
Cabeça	71.7 (21.1)	27.4 (21.1)	1.31 (7.12)
Tronco	0.84 (0.12)	92.7 (9.12)	6.51 (9.11)
Pernas	0.00 (0.00)	2.33 (6.22)	97.6 (6.24)

Analisando os resultados da atribuição de partes anatómicas, no caso em que os indivíduos se encontram isolados os resultados são bastante bons. A métrica utilizada para avaliar a segmentação de todo o corpo é muito próxima de zero. Ao nível das partes anatómicas, verifica-se que apenas a cabeça apresenta uma taxa de detecção correcta abaixo dos 90%. Isto deve-se ao facto de muitas vezes a zona da cara estar incluída num *blob* que também inclui zonas do tronco, o que acontece devido a zonas com pele do tronco estarem sobrepostas, ou muito perto, da zona da cara, o que faz com que no módulo de

Tabela 3.5: Matriz de confusão da atribuição de partes anatómicas para pessoas sobrepostas

	Cabeça	Tronco	Pernas	Erros
Cabeça	45.2 (35.8)	17.2 (20.4)	0.64 (4.52)	36.8 (42.1)
Tronco	3.73 (9.47)	62.5 (33.9)	4.92 (11.8)	28.7 (33.8)
Pernas	0.09 (0.85)	4.52 (9.11)	63.9 (34.7)	31.1 (34.7)

Tabela 3.6: Matriz de confusão da atribuição de partes anatómicas para pessoas sobrepostas, quando não são considerados os erros do módulo de gestão de sobreposição.

	Cabeça	Tronco	Pernas
Cabeça	65.5 (32.8)	33.4 (32.8)	1.07 (6.01)
Tronco	5.45 (12.1)	86.6 (21.1)	7.89 (17.5)
Pernas	0.17 (1.61)	7.51 (15.2)	92.3 (15.3)

segmentação de regiões activas seja criado apenas um *blob* que incluí a cara e os braços, como pode ser verificado na segunda coluna da Figura 3.18.

No caso da atribuição de partes anatómicas para pessoas sobrepostas os erros são bastante elevados. Em ambos os ensaios realizados, com e sem erros do módulo de gestão de sobreposição, a métrica utilizada para avaliar a segmentação de todas as partes anatómicas tem um valor bastante elevado, com cerca de um terço dos pixeis a estarem mal classificados. Ao nível dos resultados para partes anatómicas, à semelhança dos resultados apresentados para pessoas isoladas a cabeça é a parte anatómica com a taxa de detecção correcta mais baixa. No entanto, para todas as partes anatómicas a taxa de detecção correcta sofre um decrescimo. Estes resultados devem-se sobretudo a erros no módulo de gestão de sobreposição, uma vez que todos os erros aí cometidos são propagados para o módulo de atribuição de partes anatómicas. Mesmo quando a avaliação é realizada não tendo em conta os erros cometidos no módulo de gestão de sobreposição, os resultados apesar de sofrerem uma melhoria, esta não é significativa, pois o módulo de atribuição de partes anatómicas não consegue lidar com casos em que na sua entrada existe apenas uma parte de um indivíduo, como pode ser verificado na segunda coluna da Figura 3.19. À semelhança do módulo de gestão de sobreposição, verificou-se por observação experimental que ao longo de uma sobreposição os resultados da atribuição de partes anatómicas tendem a degradar-se, o que vem mais uma vez verificar que os resultados do módulo de atribuição de partes anatómicas para pessoas sobrepostas estão muito dependentes dos resultados da gestão de sobreposição.

3.6 Conclusão

Face aos resultados apresentados, podemos concluir que o sistema apresentado é efectivamente capaz de segmentar e seguir pessoas numa sequência vídeo bem como segmentar as pessoas em partes do corpo coerentes.

Os módulos de subtracção de fundo e seguimentos de regiões activas apresentam resultados muito bons, com o primeiro a apresentar erros residuais e o segundo a realizar a

Tabela 3.7: Resultados da atribuição de partes anatómicas quando avaliados com a métrica de Hammoude.

	H
Pessoas isoladas	0.05 (0.07)
Pessoas sobrepostas	0.38 (0.31)
Pessoas sobrepostas sem erros	0.24 (0.33)

tarefa proposta de forma perfeita em todas as sequências de teste.

O módulo de segmentação de regiões activas também atinge todos os requisitos propostos, no entanto, o facto de utilizar o algoritmo EM, com uma inicialização aleatória que, em geral converge para um mínimo local, faz com que os resultados dos módulos a jusante sejam dependentes da sua inicialização. Este problema poderia ter sido minimizado, se fossem realizadas várias inicializações do algoritmo EM e depois escolhida a melhor com base no critério de verosimilhança. Esta opção não foi tomada pois o custo computacional seria muito elevado, o que faria com que a utilização de um sistema deste tipo em ambiente real se tornasse impossível.

O módulo de gestão de sobreposição, é o módulo que apresenta os resultados piores. No entanto, e por observação experimental pode-se concluir que nos casos em que as pessoas não tenham grandes áreas sobrepostas e, ou, não estejam sobrepostas durante muito tempo, este módulo tem um desempenho bastante bom.

O módulo de atribuição de partes anatómicas também tem um desempenho bastante bom especialmente para pessoas isoladas, uma vez que para pessoas sobrepostas os seus resultados são muitos dependentes dos resultados do módulo de gestão de sobreposição.

O sistema tem no entanto algumas limitações devido ao facto de usar a cor como elemento distintivo, quer dos diferentes indivíduos, quer das diferentes partes anatómicas. Assim, no caso de dois indivíduos estarem vestidos da mesma maneira, ou no caso de de um individuo estar vestido todo da mesma cor os resultados dos módulos de gestão de sobreposição e atribuição de partes anatómicas sofreriam perdas significativas no seu desempenho.

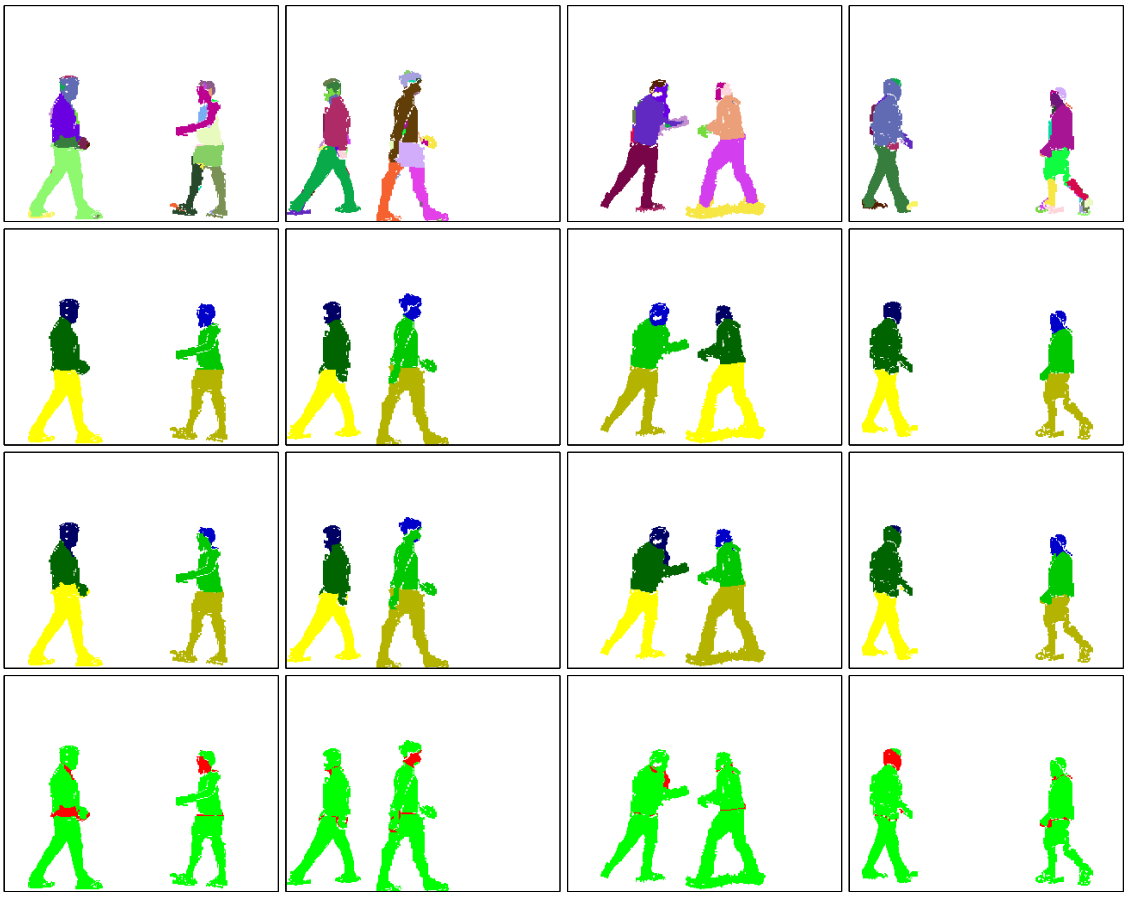


Figura 3.18: Avaliação da atribuição de partes anatómicas para pessoas isoladas: (1ª linha) entrada, (2ª linha) segmentação real, (3ª linha) segmentação estimada, (4ª linha) classificação dos pixeis, onde as cores verde e vermelho correspondem a pixeis classificados correctamente e incorrectamente, respectivamente.

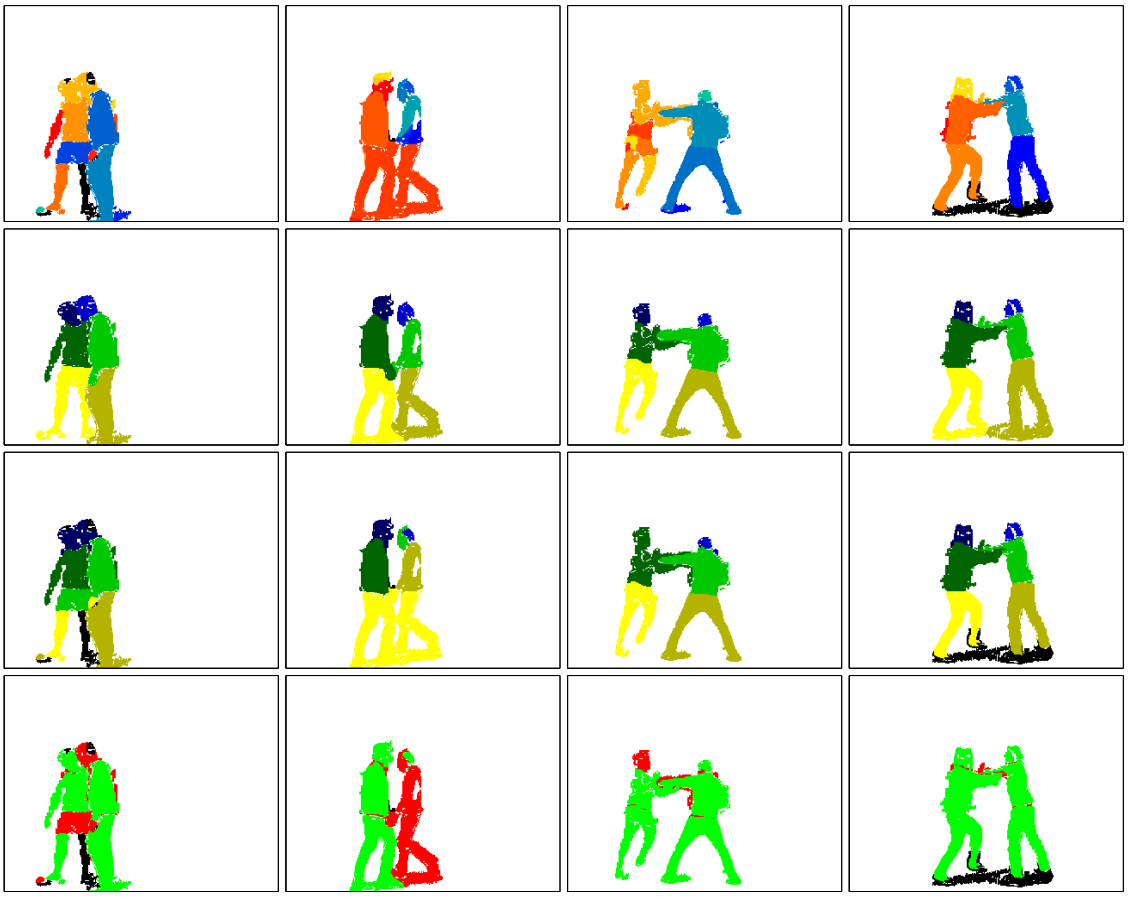


Figura 3.19: Avaliação da atribuição de partes anatómicas para pessoas sobrepostas: (1ª linha) entrada, (2ª linha) segmentação real, (3ª linha) segmentação estimada, (4ª linha) classificação dos pixeis, onde as cores verde e vermelho e preto correspondem a pixeis classificados correctamente e incorrectamente e pixeis não classificados, respectivamente.

Capítulo 4

Reconhecimento de Actividades

4.1 Introdução

Pretende-se identificar quatro tipos de actividade (abraço, cruzamento, cumprimento e luta) entre duas pessoas em vista lateral, com recurso a apenas um segundo de informação a partir do início da interacção. Decidiu-se utilizar apenas um segundo de informação, pois os seres humanos conseguem facilmente identificar uma interacção, presente numa sequência vídeo de um segundo de duração, o que mostra que existe informação suficiente sobre a interacção numa sequência de um segundo.

O início da interacção, a partir do qual se começa a recolher informação para a identificação da actividade, é escolhido quando a distância entre as coordenadas horizontais dos centroides das silhuetas dos sujeitos presentes na cena, é menor que um determinado limiar. Este limiar é normalizado pela altura média das silhuetas, por forma a ficar independente do tamanho da imagem e da profundidade a que se encontram os sujeitos. A Figura 4.1 mostra a trajectória horizontal de duas pessoas que se cumprimentam numa sequência dos dados experimentais e o intervalo de um segundo utilizado, entre a imagem 19 e a imagem 33, para caracterizar a interacção.

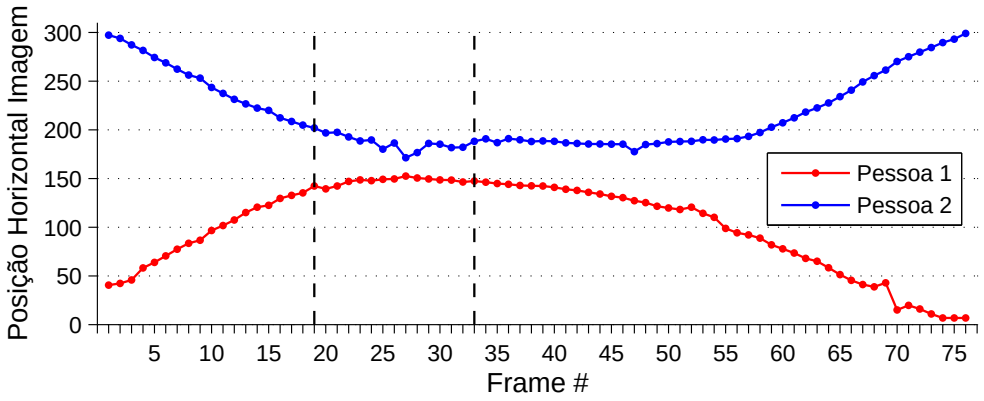


Figura 4.1: Trajectória horizontal de duas pessoas que se cumprimentam e intervalo de um segundo utilizado para caracterizar a actividade.

4.2 Características

As imagens de um vídeo podem ser representados sob a forma de uma sequência de matrizes por $F_1, F_2, ..., F_K$, onde K é o número de imagens. Assim uma sequência de um segundo, utilizada para caracterizar as actividades pode ser representada por:

$$F = \{F_1, F_2, ..., F_N\} \tag{4.1}$$

onde N é igual ao ritmo de amostragem do vídeo em análise. Cada sequência é descrita por um vector de características $x \in \mathbb{R}^n$ (n=5) em que x_i designa a i-ésimia característica.

Para cada um dos modelos estudados foram escolhidas cinco características diferentes, que compõem o vector x. As características utilizadas para cada um dos modelos foram escolhidas tendo em conta a facilidade de extracção, a robustez às oclusões e o seu poder discriminativo entre as diversas interacções que se pretendem identificar. Outra consideração que foi tida em conta na escolha das características, é que uma vez que não é utilizado nenhum método para modelar a interacção, não se pode utilizar características individuais de cada pessoa na classificação, as características devem ser descritivas das interacções em si, ou de relações entre as duas pessoas que as realizam. Para o modelo da silhueta as características extraídas procuram descrever o tipo de movimentos que as várias silhuetas presentes em cena executam. No caso do modelo anatómico procurou-se utilizar características que tirem partido do conhecimento das partes anatómicas de cada indivíduo. Segue-se uma descrição detalhada de cada característica extraída para cada um dos modelos.

4.2.1 Modelo da Silhueta

O vector x, descritivo de cada sequência video, no método do modelo da silhueta é constituído por,

- x_1: Variação de área activa.
- x_2: Uniformidade do movimento.
- x_3: Ocupação espacial.
- x_4: Média da distância entre indivíduos durante a sequência.
- x_5: Desvio padrão da distância entre indivíduos durante a sequência.

A Figura 4.2 mostra a assinatura de cada tipo de actividade, ou seja, a média do vector x para cada uma das actividades que se pretende identificar. A Figura 4.3 compara as assinaturas de cada tipo de actividade. Descreve-se de seguida o processo de cálculo de cada uma das características que compõem o vector x.

Variação de Área Activa

A variação de área activa é calculada fazendo a média da variação da sequência video. A variação é dada pela percentagem de pixeis que se encontram activos em F_k que não se encontravam activos em F_{k-1}. Se chamarmos ${}^1F_k$ ao conjunto de pixeis activos em F_k a taxa de variação é dada por

$$x_1 = \frac{1}{N} \sum_{k=1}^{N} \frac{\#{}^1F_k - \#({}^1F_k \cap {}^1F_{k-1})}{\#{}^1F_k} \tag{4.2}$$

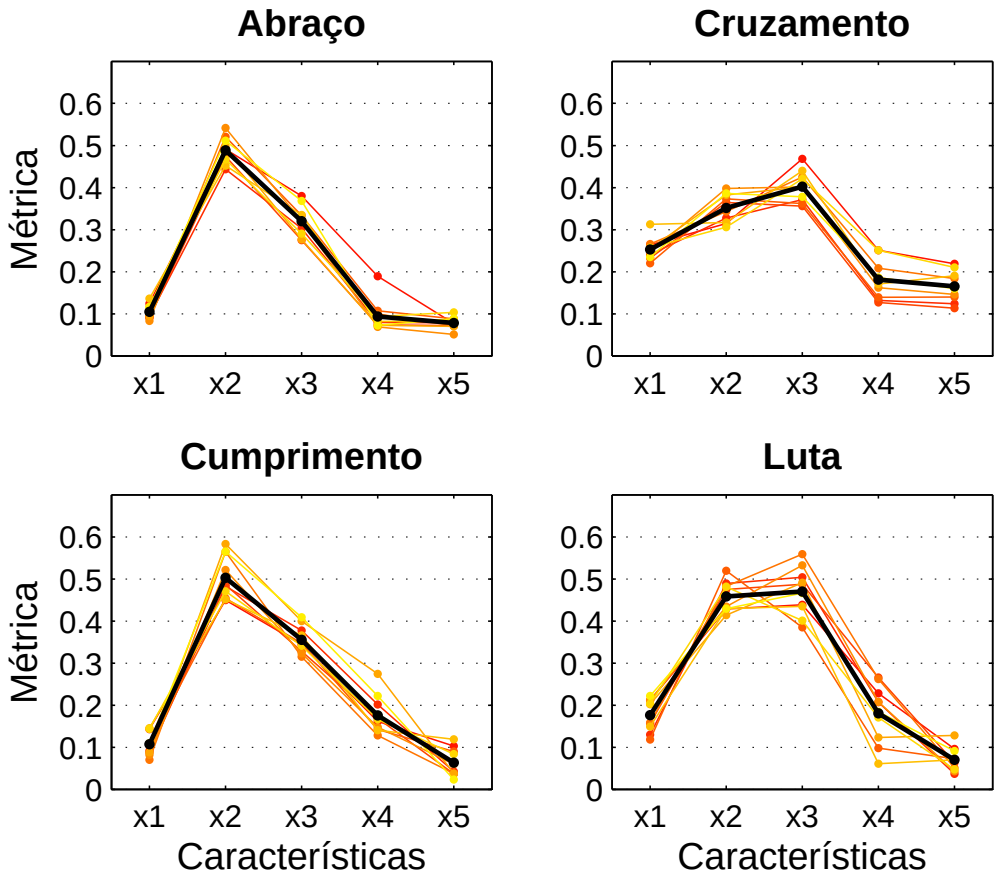


Figura 4.2: Assinatura de cada uma das actividades que se pretende reconhecer.

A variação de área activa reflecte, a velocidade do movimento efectuado durante o segundo em análise. Assim uma taxa de variação de 1 indica que houve movimento rápido, enquanto um valor de zero reflecte ausência de movimento.

Uniformidade do Movimento

Para calcular a uniformidade de movimento é construída uma representação MEI (Motion Energy Image), à semelhança da proposta em [38], onde são agregadas as regiões activas de todas as imagens que constituem a sequência F. A Figura 4.4 ilustra a representação MEI de uma sequência extraída dos dados experimentais. A representação MEI pode ser interpretada como uma matriz J dada por,

$$J = \frac{1}{N}\sum_{k=1}^{N} {}^{1}F_k \tag{4.3}$$

A uniformidade de movimento é então dada pelo desvio padrão de todos os elementos da matriz J,

$$x_2 = \sqrt{\frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}(J_{ij} - \bar{J})^2} \tag{4.4}$$

onde $\bar{J}$ é a média de todos os elementos da matriz V dada por,

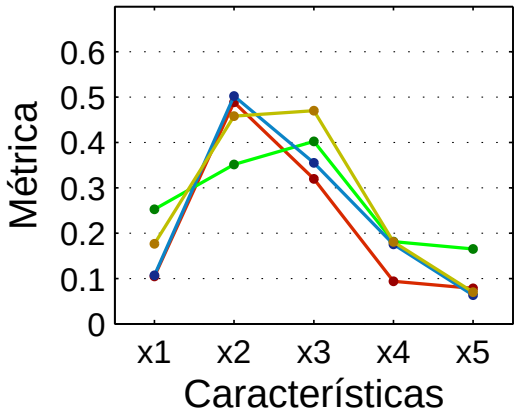


Figura 4.3: Comparação entre as assinaturas de cada uma das actividades que se pretende reconhecer, onde as cores vermelho, verde, azul e amarelo correspondem às actividades abraço, cruzamento, cumprimento e luta respectivamente.

$$\bar{V} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij} \tag{4.5}$$

A uniformidade de movimento tem valores elevados, quando o movimento durante a sequência é uniforme e valores baixos quando existem mudanças de velocidade no movimento efectuado, como por exemplo dois indivíduos aproximarem-se e depois pararem.

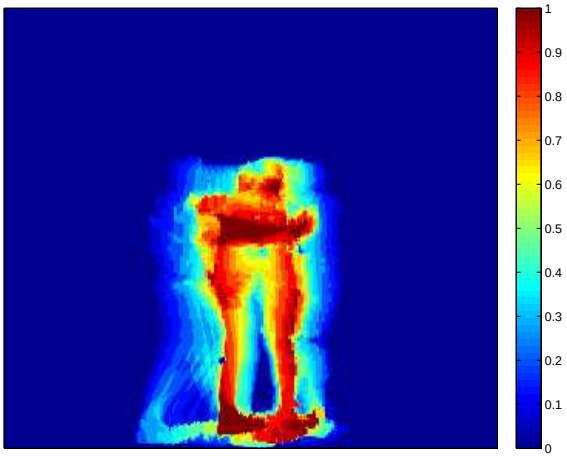


Figura 4.4: Representação MEI de uma sequência v extraída dos dados experimentais

Ocupação Espacial

Para cada imagem $F_k \in F$ é extraída a *bounding box* que contém todas as silhuetas presentes em cena. A ocupação espacial é então dada pela média da largura da *bounding box*, normalizada pela altura da mesma, em todas as imagens $F_k \in F$. Esta característica descreve o espaço ocupado pelos dois indivíduos ao longo da sequência.

Distância entre indivíduos

A distância entre indivíduos é calculada com recurso a uma mistura de Gaussianas unidimensionais treinada de forma semelhante à apresentada na secção 3.4.2. Uma vez treinadas as Gaussianas, a distância entre indivíduos é dada pelo absoluto da diferença das médias das duas Gaussianas estimadas. Esta distância é calculada em todas as imagens da sequência vídeo de um segundo em análise, de seguida são calculadas a média e o desvio padrão da distância entre indivíduos ao longo da sequência e são estas medidas que são utilizadas como descritores da interacção.

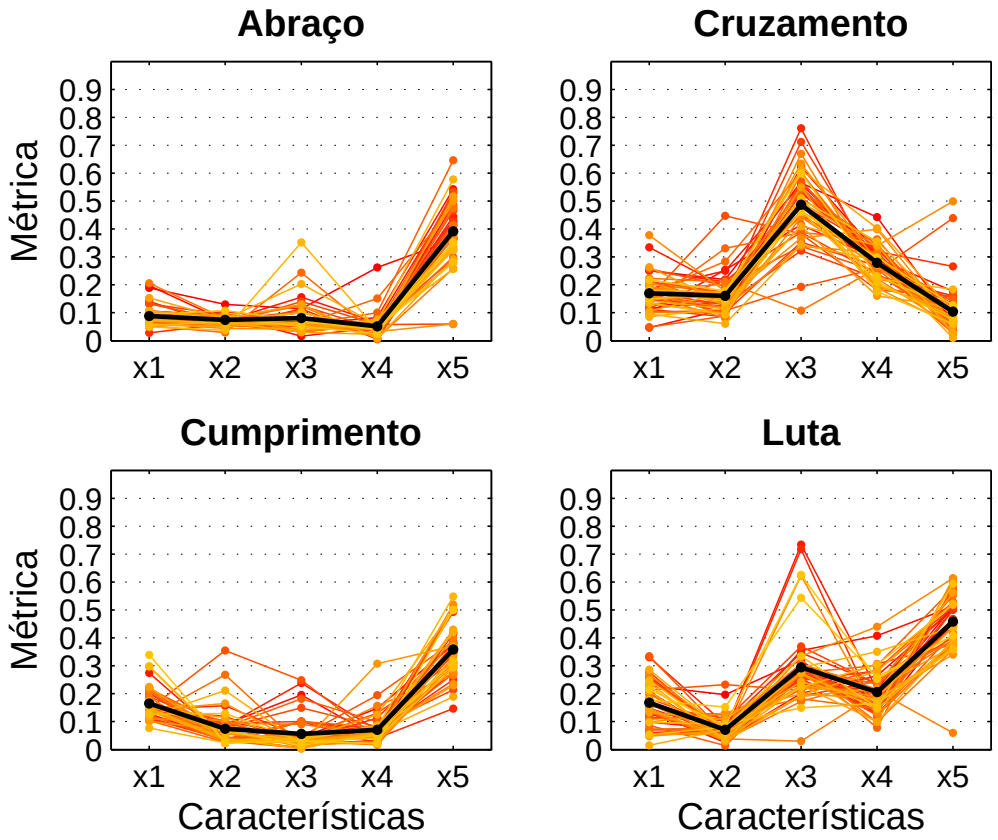


Figura 4.5: Assinatura de cada uma das actividades que se pretende reconhecer.

4.2.2 Modelo Anatómico

O vector x, descritivo de cada sequência, no método do modelo da silhueta é constituído por,

- x_1: Média da distância entre as cabeças dos indivíduos durante a sequência.
- x_2: Desvio padrão da distância entre as cabeças dos indivíduos durante a sequência.
- x_3: Média da distância entre as mãos dos indivíduos durante a sequência.
- x_4: Desvio padrão da distância entre as mãos dos indivíduos durante a sequência.
- x_5: Diferença entre o ângulo do braço mais afastado do tronco de cada indivíduo no início da sequência v.

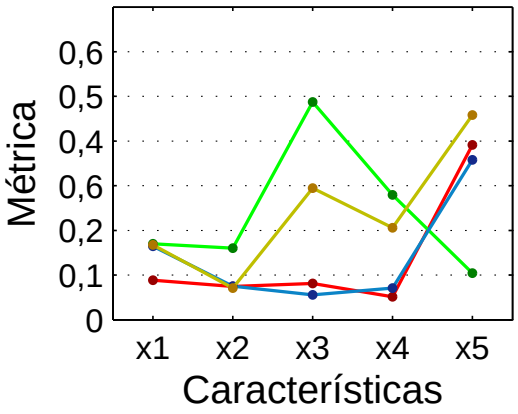


Figura 4.6: Comparação entre as assinaturas de cada uma das actividades que se pretende reconhecer, onde as cores vermelho, verde, azul e amarelo correspondem às actividades abraço, cruzamento, cumprimento e luta respectivamente.

A Figura 4.5 mostra a assinatura de cada tipo de actividade, ou seja a composição típica do vector x para cada uma das actividades que se pretende identificar. A Figura 4.6 compara as assinaturas de cada tipo de actividade. Descreve-se de seguida o processo de cálculo de cada uma das características que compõem o vector x.

Distância entre cabeças

A distância entre os centróides das cabeças é dada pela norma dos centróides das cabeças de cada indivíduo, normalizada pela altura da *bounding box* que contém todos os indivíduos presentes na imagem em análise. Esta distância é calculada em todas as imagens que compõem a sequência F, sendo depois calculada a respectiva média e desvio padrão, que são utilizados como descritores da actividade.

Distância entre mãos

Para cada indivíduo é estimada a posição da mão que se encontra mais afastada do corpo. A posição da mão mais afastada do corpo é dada pelo ponto de curvatura máxima do *convex hull* do tronco, que coincide com um *blob* marcado como pele. No caso de existirem vários pontos de curvatura máxima coincidentes com um *blob* marcado como pele, é escolhido aquele que estiver mais afastado horizontalmente do centróide do corpo. A Figura 4.7 mostra um exemplo de estimação da posição da mão mais afastada do tronco, onde os pontos A e B são as posições estimadas.

A distância entre mãos é dada pela norma da posição da mão mais afastada do corpo de cada indivíduo, normalizada pela altura da *bounding box* que contém todos os indivíduos presentes na imagem em análise. Esta distância é calculada em todas as imagens $F_k \in K$, sendo depois calculada a respectiva média e desvio padrão, que são utilizados como descritores da actividade.

Diferença entre ângulos do braço

Para cada pessoa é calculado o ângulo entre o braço mais afastado e o tronco. Este ângulo é definido como o ângulo entre uma recta vertical que atravessa o ombro, e uma recta que

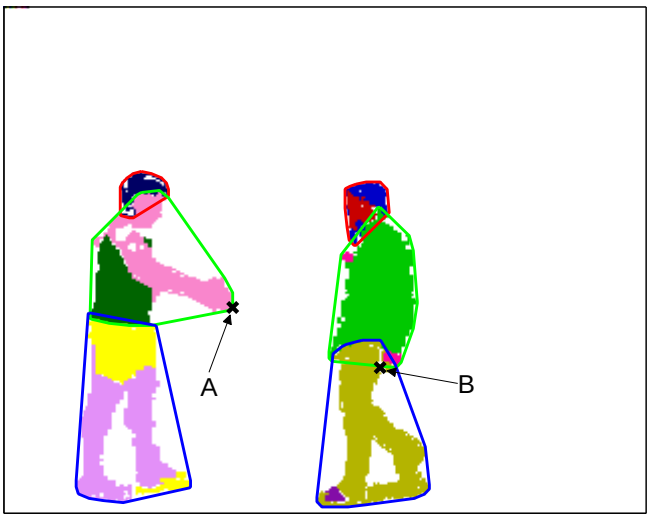


Figura 4.7: *Convex Hull* de cada parte do corpo. Os pontos de curvatura máxima A e B são detectados como possíveis candidatos para posição das mãos

une o ombro à mão mais afastada do tronco. A posição do ombro é dada pelo ponto de curvatura máxima do *convex hull* que contém o tronco, que está mais perto da cabeça. De seguida é calculada a diferença entre os ângulos dos braços de cada indivíduo e normalizada por 360^{o}. Este valor é calculado apenas na primeira imagem da sequência F e é utilizado como descritor da sequência. A Figura 4.8 ilustra um exemplo do cálculo do ângulo do braço.

4.3 Classificador

O classificador utilizado para identificar as diversas interacções, é um classificador de K vizinhos mais próximos baseado na distância Euclideana. Escolheu-se utilizar este classificador pois é fácil de treinar, ou seja, não é necessário ter um conjunto de dados de treino extenso, para conseguir que o classificador tenha um desempenho consistente. Outra característica importante deste classificador que foi tida em conta, é a sua facilidade de implementação e o facto de não ser necessário ter conhecimento *a priori* da distribuição dos dados.

4.4 Resultados Experimentais

Para cada um dos modelos em estudo o desempenho do classificador foi avaliado pelo método *leave one out*, ou seja o conjunto de dados utilizado para treinar o classificador é constituído por todas as sequências, menos a sequência usada para teste. No caso do método da silhueta foi utilizado um classificador de K vizinhos mais próximos que utiliza os 3 vizinhos mais próximos de entre 39 para classificar a actividade. No caso do modelo anatómico, como a extracção de características depende da inicialização do algoritmo EM, as sequências de vídeo foram processadas com cinco inicializações diferentes e todas as sequência foram utilizadas. Assim neste caso consideraram-se 200 sequências, cinco inicializações das 40 sequências do conjunto de dados experimentais. O classificador, no caso do método do modelo anatómico utiliza os 15 vizinhos mais próximos de entre 195

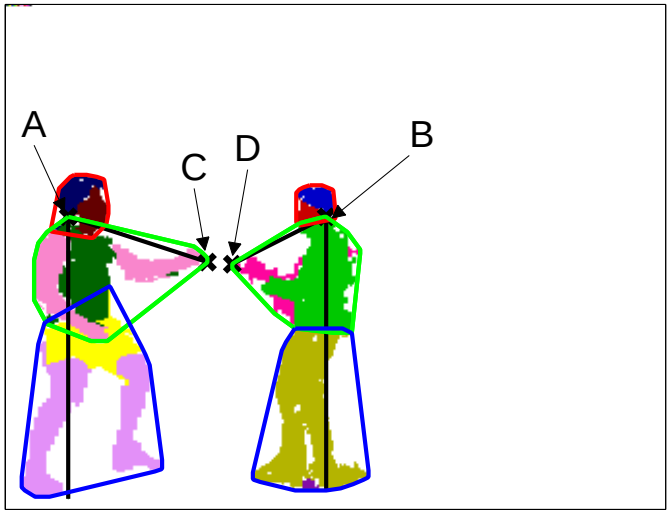


Figura 4.8: Exemplo de cálculo do ângulo do braço. Os pontos A e B representam posições estimadas para os ombros, e os pontos C e D representam posições estimadas para as mãos.

para classificar a actividade. A Tabela 4.1 apresenta os resultados do reconhecimento de actividades para o modelo da silhueta, enquanto a tabela 4.2 apresenta os resultados obtidos com o modelo anatómico.

Tabela 4.1: Resultados do reconhecimento de actividade utilizando o método do modelo da silhueta.

	abraço	cruzamento	cumprimento	luta
abraço	90	0	10	0
cruzamento	0	100	0	0
cumprimento	0	0	100	0
luta	20	0	0	80

4.5 Conclusão

No método do modelo da silhueta a taxa global de sucesso é de 92.5%. O cruzamento e o cumprimento são identificados com 100% de sucesso, enquanto a taxa de sucesso para o abraço é de 90% e para a luta é de 80%. A luta é a actividade, que como se esperava, tem um taxa de identificação mais baixa, pois os indivíduos durante este tipo de interacção têm um conjunto de movimentos erráticos, com as características extraídas a possuirem uma maior variância, como pode ser observado no Figura 4.2, o que torna mais díficil a sua caracterização. Todas as sequências da actividade abraço que não foram correctamente classificadas, foram classificadas como cumprimento, este resultado era esperado, pois analisando a Figura 4.3 pode-se verificar que as assinaturas destas duas interacções são muito semelhantes.

Tabela 4.2: Resultados do reconhecimento de actividades utilizando método do modelo anatómico.

	abraço	cruzamento	cumprimento	luta
abraço	86	0	12	2
cruzamento	0	96	0	4
cumprimento	12	2	84	2
luta	2	4	6	88

No método do modelo anatómico a taxa global de sucesso é de 90.8%. O cruzamento é a actividade com maior taxa de sucesso, 96% seguido da luta 88%, abraço 86% e finalmente o cumprimento 84%. O abraço e o cumprimento são as actividades com pior taxa de sucesso uma vez que o classificador, classifica alguma vezes abraços como sendo cumprimentos e cumprimentos como sendo abraços. Isto deve-se ao facto de como pode ser visto na Figura 4.6 as assinaturas destas actividades são muito semelhantes. A luta apresenta uma taxa de sucesso maior quando comparada com a taxa de sucesso no método da silhueta, isto porque as características utilizadas no método do modelo anatómico são melhores descritores desta interacção.

Globalmente a actividade com a maior taxa de sucesso é o cruzamento. Uma vez que esta actividade é uma não interacção, é a actividade mais facilmente descriminada.

Capítulo 5

Conclusão

Nesta tese foi implementado e testado um sistema automático de reconhecimento de interacções entre duas pessoas, e comparou-se a utilização de dois modelos do corpo humano para esse fim. Foram realizadas algumas hipóteses com vista à simplificação do problema, e face a essas hipóteses o sistema implementado mostrou ser efectivamente capaz de identificar interacções utilizando ambos os modelos do corpo humano testados.

É proposto um método para resolver oclusões entre pessoas que mostrou ser robusto e fiável, ao não apresentar nenhum erro quando testado com os dados experimentais recolhidos. No entanto, este método assume que as pessoas sobrepostas utilizam pelo menos uma peça de roupa, no tronco ou pernas, de cor diferente entre si, situação que é verdadeira nos dados experimentais mas que num cenário de vídeo vigilância nem sempre se verifica. Este método pode ser bastante útil em cenários onde a resolução seja baixa e não seja possível extrair outras características que permitam identificar a identidade de cada pessoa após uma oclusão.

É proposto um método para segmentar duas pessoas que se encontrem sobrepostas. Este método apresenta alguns erros, especialmente quando existem grandes regiões de cada indivíduo ocludidas. No entanto quando as pessoas sobrepostas apenas têm partes pequenas do corpo ocludidas o método mostrou ser robusto. Os erros devem-se sobretudo a serem criados *blobs* no módulo de segmentação de regiões activas que contêm partes das duas pessoas sobrepostas, o que impossibilita a sua correcta classificação.

São propostas cinco características descritivas de uma sequência vídeo de um segundo quando é utilizado um modelo da silhueta para representar o corpo humano. Estas características mostraram ter poder discriminativo suficiente para permitir a classificação das interacções que esta tese se propunha identificar. A extracção das características é fácil, robusta e de baixo custo computacional. As características extraídas são no entanto dependentes da pose das pessoas em relação à camera. O seu poder discriminativo pode apenas ser garantido, tendo em conta as hipóteses assumidas.

São propostas cinco características descritivas de uma sequência vídeo de um segundo quando é utilizado um modelo anatómico para representar o corpo humano. Estas características mostraram ter poder discriminativo suficiente para permitir a classificação das interacções que esta tese se propunha identificar. A extracção das características é fácil e de baixo custo computacional, no entanto a robustez da sua extracção está dependente da correcta estimação do modelo anatómico. À semelhança das características extraídas para o modelo da silhueta, estas são dependentes da pose das pessoas em relação à camera e o seu poder discriminativo pode apenas ser garantido, tendo em conta as hipóteses assumidas.

Comparando os dois modelos testados nesta tese, e a sua utilização num sistema de

reconhecimento automático de actividades, ambos apresentam vantagens e desvantagens. O modelo da silhueta obteve os melhores resultados, com uma taxa de classificações correctas de 92.5%, e é o modelo cuja extracção é mais robusta, mais fácil de implementar e de menor custo computacional. No entanto, quando duas pessoas estão sobrepostas este modelo não possibilita a extracção de características que sejam descritivas de cada um dos indíviduos, as características extraídas devem ser descritivas da interacção em si. Outro problema deste modelo é que algumas das características que foram extraídas descrevem o movimento global da cena, o que impossibilita o seu uso no caso em que existam mais pessoas em cena que não estejam a interagir. O modelo anatómico obteve resultados semelhantes aos do modelo da silhueta com uma taxa de classificações correctas de 90.8%. Este modelo descreve muito melhor as pessoas que realizam as interacções, permitindo extrair características de cada indivíduo mesmo quando estes se encontram sobrepostos. No entanto a construção do modelo anatómico é complexa e de elevado custo computacional, e necessita que as pessoas estejam relativamente perto da camera ou que a resolução das imagens capturadas seja elevada. Face às hipóteses que foram assumidas nesta tese, o modelo da silhueta assume-se como a abordagem mais eficaz, uma vez que apresentou os melhores resultados e é uma abordagem muito mais simples e robusta. Por outro lado, o modelo anatómico tem mais potencial para ser utilizado num ambiente de vídeo vigilância real, em que existam mais pessoas em cena e onde a pose das pessoas não seja conhecida, uma vez que este é mais descritivo de cada indivíduo e permite a segmentação de pessoas ocludidas.

Capítulo 6

Trabalho Futuro

O sistema automático de reconhecimento de interacções implementado nesta tese apenas funciona num ambiente controlado. Para se ter um sistema capaz de funcionar num ambiente de vídeo vigilância real teriam que se eliminar algumas das hipóteses assumidas nesta tese.

Uma das limitações do sistema consiste no facto de o módulo de subtracção de fundo ser apenas capaz de lidar com casos em que o fundo é estático, situação que num ambiente real nem sempre é verdadeira. Era então necessário implementar um método para segmentar as regiões activas de cada imagem que fosse capaz de lidar com fundos dinâmicos.

Outra restrição do trabalho proposto reside na capacidade deste apenas conseguir lidar com casos em que as pessoas realizam as interacções em vista lateral. Seria interessante implementar um método para estimar a pose de cada sujeito em relação à camera, de forma a se poder extrair características distintas para cada pose e tornar o sistema multi-vista.

Outra direcção a tomar para tornar o sistema capaz de funcionar com mais de duas pessoas em cena seria a de extrair características descritivas de cada pessoa ao invés de características descritivas da interacção. O classificador utilizado teria então que ser mais sofisticado, porque tinha não só de classificar as interacções, mas identificar quando estas começam e que sujeitos as realizam.

Ainda um outro caminho que seria interessante explorar, seria o de integrar outros sensores no sistema, tais como sensores de profundidade ou microfones. Estas novas ferramentas iriam trazer todo um novo potencial para conseguir descrever o que as pessoas estão a fazer em cena com muito mais detalhe e rigor. O futuro dos sistemas de vigilância inteligentes passa pela implementação de sistemas multi-sensoriais, com algum nível de redundância para que possam efectivamente cumprir os objectivos a que se propõem de forma robusta e fiável.

Bibliografia

[1] K. Jia and D.-Y. Yeung, "Human action recognition using local spatio-temporal discriminant embedding," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1 –8, june 2008.

[2] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 984 – 989 vol. 1, june 2005.

[3] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–819 – II–826 Vol.2, jun. 2004.

[4] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa, "Shape activity: a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection," *Image Processing, IEEE Transactions on*, vol. 14, pp. 1603 –1616, oct. 2005.

[5] S. Park and J. Aggarwal, "Semantic-level understanding of human actions and interactions using event hierarchy," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, p. 12, june 2004.

[6] Y. Du, F. Chen, and W. Xu, "Human interaction representation and recognition through motion decomposition," *Signal Processing Letters, IEEE*, vol. 14, pp. 952 –955, dec. 2007.

[7] H.-I. Suk, B.-K. Sin, and S.-W. Lee, "Analyzing human interactions with a network of dynamic probabilistic models," in *Applications of Computer Vision (WACV), 2009 Workshop on*, pp. 1 –6, dec. 2009.

[8] J. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[9] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.

[10] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, pp. 1151 – 1163, jul. 2002.

[11] M. Harville, G. Gordon, and J. Woodfill, "Foreground segmentation using adaptive mixture models in color and depth," *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pp. 3 –11, 2001.

[12] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz, "Hmm-based human motion recognition with optical flow data," *Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on*, pp. 425 –430, dec. 2009.

[13] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical flow analysis for lip images," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 36, no. 2- 3, pp. 117 –124, 2004.

[14] S. Beauchemin and J. Barron, "The computation of optical flow," 1995.

[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 780 –785, jul. 1997.

[16] S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video," *Proc Elect Imaging: Visual Comm Image Proce 2004 (Part One)*, pp. 881 –892, 2004.

[17] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2 -3, pp. 107 –123, 2005.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65 – 72, oct. 2005.

[19] C. Veenman, M. Reinders, and E. Backer, "Resolving motion correspondence for densely moving points," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 54 –72, Jan. 2001.

[20] Y. Bar-Shalom and T. Fortmann, "Tracking and data association," *Academic Press Inc*, 1988.

[21] I. Cox and S. Hingorani, "An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, pp. 138 –150, feb. 1996.

[22] R. Hess and A. Fern, "Discriminatively trained particle filters for complex multi-object tracking," *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 240 –247, jun. 2009.

[23] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, pp. 564 – 577, may. 2003.

[24] A. Jepson, D. Fleet, and T. El-Maraghi, "Robust online appearance models for visual tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, pp. 1296 – 1311, oct. 2003.

[25] J. Shi and C. Tomasi, "Good features to track," *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, pp. 593 –600, jun. 1994.

[26] M. Liu, C. Wu, and Y. Zhang, "Multi-resolution optical flow tracking algorithm based on multi-scale harris corner points feature," *Control and Decision Conference, 2008. CCDC 2008. Chinese*, pp. 5287 –5291, jul. 2008.

[27] M. Black and A. Jepson, "Recognizing temporal trajectories using the condensation algorithm," *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pp. 16 –21, apr. 1998.

[28] S. Avidan, "Support vector tracking," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–184 – I–191 vol.1, 2001.

[29] B. Li, R. Chellappa, Q. Zheng, and S. Der, "Model-based temporal object verification using video," *Image Processing, IEEE Transactions on*, vol. 10, pp. 897 –908, jun. 2001.

[30] J. Kang, I. Cohen, and G. Medioni, "Object reacquisition using invariant appearance model," *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, pp. 759 – 762, aug. 2004.

[31] K. Sato and J. K. Aggarwal, "Temporal spatio-velocity transform and its application to tracking and interaction," *Comput. Vision Image Understanding*, vol. 96, pp. 100 –128, 2004.

[32] Y. Chen, Y. Rui, and T. Huang, "Jpdaf based hmm for real-time contour tracking," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–543 – I–550 vol.1, 2001.

[33] N. Vaswani, Y. Rathi, A. Yezzi, and A. Tannenbaum, "Deform pf-mt: Particle filter with mode tracker for tracking nonaffine contour deformations," *Image Processing, IEEE Transactions on*, vol. 19, pp. 841 –857, apr. 2010.

[34] M. Yokoyama and T. Poggio, "A contour-based moving object detection and tracking," *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 271 – 276, oct. 2005.

[35] K. Takaya, "Tracking a video object with the active contour (snake) predicted by the optical flow," *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*, pp. 000369 –000372, may. 2008.

[36] A. Yilmaz, X. Li, and M. Shah, "Contour-based object tracking with occlusion handling in video acquired using mobile cameras," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 1531 –1536, nov. 2004.

[37] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, pp. 1473 –1488, nov. 2008.

[38] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 257 –267, mar. 2001.

[39] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, pp. 179 –187, february 1962.

[40] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1395 –1402, oct. 2005.

[41] O. Chomat and J. Crowley, "Probabilistic recognition of activity using local appearance," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2, pp. 2 vol. (xxiii+637+663), 1999.

[42] L. Zelnik-Manor and M. Irani, "Event-based analysis of video," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–123 – II–130 vol.2, 2001.

[43] J. C. Niebles, H. Wang, and L. Fei Fei, "Unsupervised learning of human actions categories using spatial-temporal words," *British Machine Vision Conference*, 2006.

[44] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," *Internacional Conference on Pattern Recognition*, pp. 32 –36, 2004.

[45] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17 –31, 2007.

[46] S.-F. Wong, T.-K. Kim, and R. Cipolla, "Learning motion categories using both semantic and structural information," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –6, june 2007.

[47] J. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –8, june 2007.

[48] E. Shechtman and M. Irani, "Space-time behavior based correlation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 405 – 412 vol. 1, june 2005.

[49] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 166 – 173 Vol. 1, oct. 2005.

[50] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," in *In 7th Int. Workshop on Visual Surveillance*, pp. 166 –173, 2007.

[51] M. A. O. Vasilescu, "Human motion signatures: Analysis, synthesis, recognition," in *In proceedinds of the international conference on pattern recognition (ICPR 02)*, pp. 456 –460, 2002.

[52] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –8, june 2007.

[53] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank svm," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –6, june 2007.

[54] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pp. 379 –385, jun 1992.

[55] J. Schlenzig, E. Hunter, and R. Jain, "Recursive identification of gesture inputs using hidden markov models," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp. 187 –194, dec 1994.

[56] A. Wilson and A. Bobick, "Learning visual behavior for gesture analysis," in *Computer Vision, 1995. Proceedings., International Symposium on*, pp. 229 –234, Nov. 1995.

[57] T. Starner and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," in *Computer Vision, 1995. Proceedings., International Symposium on*, pp. 265 –270, Nov. 1995.

[58] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 994 –999, jun 1997.

[59] Z. Moghaddam and M. Piccardi, "Deterministic initialization of hidden markov models for human action recognition," in *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.*, pp. 188 –195, dec. 2009.

[60] Z. Moghaddam and M. Piccardi, "Histogram-based training initialisation of hidden markov models for human action recognition," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 256 –261, sept. 2010.

[61] M. Mendoza, N. de la Blanca, and M. Marin-Jimenez, "Pohmm-based human action recognition," in *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, pp. 85 –88, may 2009.

[62] N. Li and D. Xu, "Action recognition using weighted three-state hidden markov model," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pp. 1428 –1431, oct. 2008.

[63] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto, "Recognition of human gaits," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–52 – II–57 vol.2, 2001.

[64] M. Mazzaro, M. Sznaier, and O. Camps, "A model (in)validation approach to gait classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1820 –1825, nov. 2005.

[65] N. Cuntoor and R. Chellappa, "Epitomic representation of human activities," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, june 2007.

[66] R. Vidal and P. Favaro, "Dynamicboost: Boosting time series generated by dynamical systems," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1 –6, oct. 2007.

[67] A. Bissacco and S. Soatto, "On the blind classification of time series," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1 –7, june 2007.

[68] S. M. Oh, J. Rehg, T. Balch, and F. Dellaert, "Learning and inference in parametric switching linear dynamic systems," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, pp. 1161 –1168 Vol. 2, oct. 2005.

[69] V. Pavlovic and J. Rehg, "Impact of dynamic model learning on classification of human motion," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 788 –795 vol.1, 2000.

[70] L. Blackmore, S. Rajamanoharan, and B. Williams, "Active estimation for switching linear dynamic systems," in *Decision and Control, 2006 45th IEEE Conference on*, pp. 137 –144, dec. 2006.

[71] S. Park and J. Aggarwal, "Segmentation and tracking of interacting human body parts under occlusion and shadowing," in *Motion and Video Computing, 2002. Proceedings. Workshop on*, pp. 105 – 111, dec. 2002.

[72] T. Yang, Q. Pan, J. Li, and S. Li, "Real-time multiple objects tracking with occlusion handling in dynamic scenes," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 970 – 975 vol. 1, jun. 2005.

[73] H. Kuhn, "The hungarian method for solving the assignment problem," *Naval Research Logistics Quart.*, vol. 2, pp. 83 – 97, 1955.

[74] V. V. Vassili, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," *in Proc. Graphicon-2003*, pp. 85– 92, 2003.

[75] R. Duda, P. Hart, and E. Stork, *Pattern Classification second ed.*, ch. 10, pp. 517 –583. Wiley, New York, 2001.

[76] L. Salgado, N. Garcia, J. Menendez, and E. Rendon, "Efficient image segmentation for region-based motion estimation and compensation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, pp. 1029 –1039, oct. 2000.

[77] Y. Bar-Shalom and W. Blair, *Multitarget-multisensor tracking: applications and advances*, vol. 3, pp. 199 –231. Norwood, MA, 2000.

[78] A. Hammoude, *Computer-assisted endocardial border identification from a sequence of two-dimensional echocardiographic images.* PhD thesis, University of Washington, Seattle, WA, 1988.