Detecção e rastreamento de múltiplos objetos em condição de oclusão severa por meio de integração de suporte sob restrição homográfica

Thiago Teixeira Santos

Tese apresentada

ao

Instituto de Matemática e Estatística

da

Universidade de São Paulo

para

obtenção do grau

de

Doutor em Ciências

Programa: **Ciência da Computação**

Orientador: **Profª Dr. Carlos Hitoshi Morimoto**

São Paulo – 2009

À Verônica dedico esta nossa algoz

Resumo

O presente trabalho propõe métodos para localizar e rastrear indivíduos combinando evidência oriunda de múltiplas câmeras, através da restrição homográfica induzida pelo plano do solo. Os procedimentos propostos utilizam um subtrator de fundo para definir quais pixeis pertencem aos objetos de interesse. Esses pixeis são empregados como evidência da localização de pessoas no plano de referência. Os algoritmos propostos computam a quantidade de *suporte*, que corresponde à "massa" observada acima de cada pixel. Pixeis que correspondem às localizações no solo onde se encontram os indivíduos irão apresentar maior suporte. Esse suporte é normalizado para compensar efeitos de perspectiva e acumulado no plano de referência para todas as câmeras observadas. A detecção de pessoas no plano do solo torna-se o problema de busca por regiões de máximos locais no acumulador. Falsos-positivos são filtrados através de uma avaliação de consistência entre os candidatos encontrados. Os candidatos remanescentes são rastreados através de Filtros de Kalman e um modelo de aparência multicâmera. Resultados experimentais a partir de dados provenientes das bases públicas PETS 2006 e PETS 2009 demonstram a eficácia dos métodos em presença de oclusão parcial ou total entre os indivíduos.

Abstract

This paper proposes a method to locate and track people by combining evidence from multiple cameras using the homography constraint. The proposed method use foreground pixels from simple background subtraction to compute evidence of the location of people on a reference ground plane. The algorithm computes the amount of support that basically corresponds to the "foreground mass" above each pixel. Therefore, pixels that correspond to ground points have more support. The support is normalized to compensate for perspective effects and accumulated on the reference plane for all camera views. The detection of people on the reference plane becomes a search for regions of local maxima in the accumulator. Many false positives are filtered by checking the visibility consistency of the detected candidates against all camera views. The remaining candidates are tracked using Kalman filters and appearance models. Experimental results using challenging data from PETS'06 and PETS'09 show good performance of the method in the presence of severe occlusion.

ÍNDICE

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

INTRODUÇÃO

Detectar e rastrear indivíduos em ambientes onde há diversas pessoas é um problema não trivial, sobretudo se consideradas as situações de oclusão entre esses indivíduos. O emprego de múltiplas câmeras em diferentes ângulos de visão é um recurso comumente empregado na resolução desse problema. O objetivo do presente trabalho é estimar a posição e a trajetória de cada pessoa, integrando de forma escalável a informação de cada câmera disponível.

1.1 Localização de pessoas sob oclusão mútua

A construção de um sistema capaz de encontrar a posição de cada indivíduo de acordo com uma planta do local de interesse, determinando a trajetória de cada indivíduo ao longo do tempo, como ilustrado na Figura 1.1.1, é o objetivo da desta tese.

Portanto, este trabalho visa compreender como a informação oriunda de n câmeras pode ser integrada e como os resultados obtidos podem ser refinados (i) à medida que mais câmeras se tornam disponíveis e (ii) conforme mais parâmetros da cada câmera são conhecidos.

Recentemente, diversos trabalhos têm sugerido abordagens simples, baseadas em *homografias entre planos*, que podem ser usadas para relacionar informação proveniente de uma rede esparsa de câmeras. A restrição imposta pela homografia estabelece que múltiplas projeções do eixo principal de um objeto, quando mapeadas para um plano de referência, intersectam-se na posição do objeto nesse plano. Essa será a propriedade empregada na

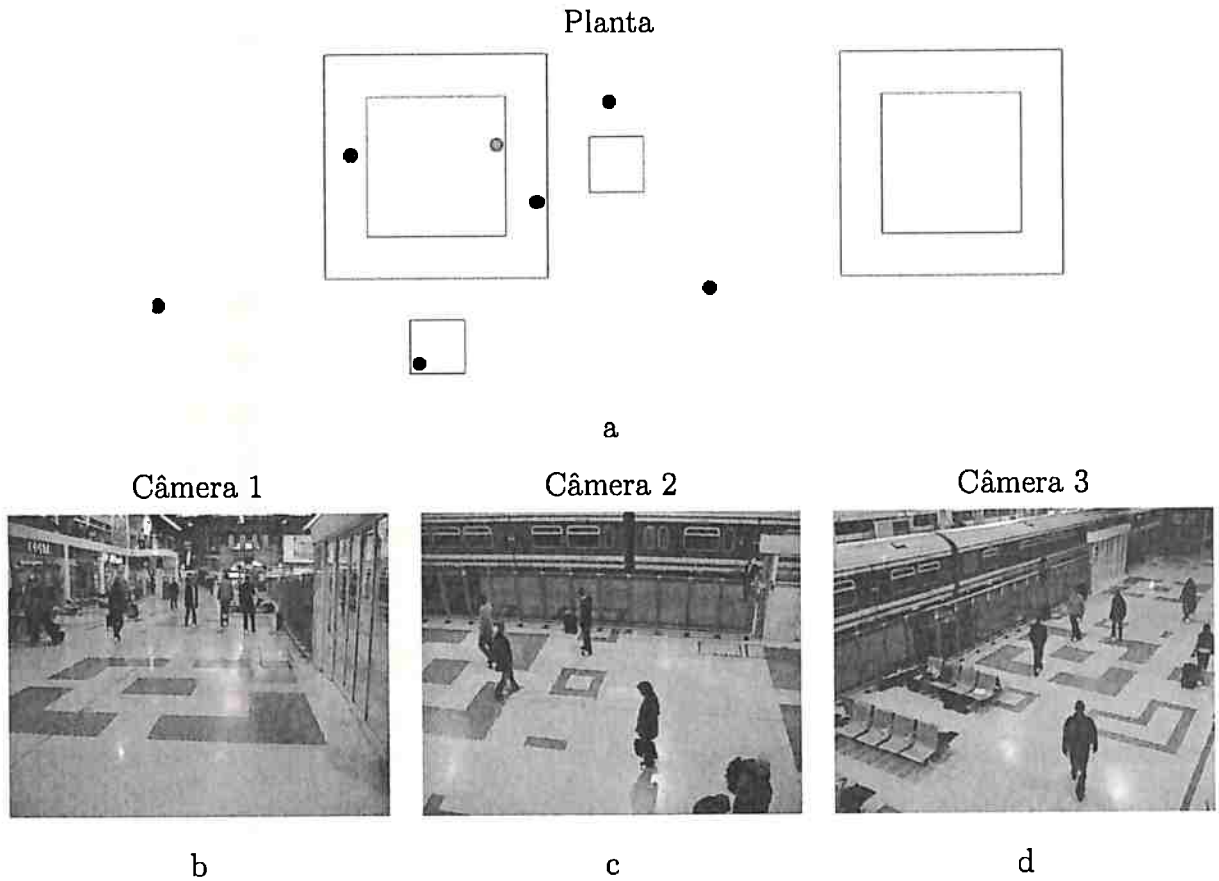


Figura 1.1.1: **Localização utilizando múltiplas câmeras.** Cena gravada em Victoria Station, Londres, proveniente da base de dados do PETS 2006 [48]. As diversas pessoas presentes ocludem umas às outras, total ou parcialmente. O objetivo deste trabalho é determinar a posição de cada indivíduo, tanto na planta do local (a) como nas imagens de entrada (b), (c) e (d), utilizando a informação de todas as câmeras disponíveis.

integração de múltiplos sensores neste estudo.

Em ambientes com várias pessoas, a oclusão entre indivíduos, total ou parcial, é frequente. Em lugares como estações de trem e aeroportos, é comum que pessoas andem em pequenos grupos a maior parte do tempo, resultando em oclusão em todas ou na maior parte das câmeras de vigilância. Nesses cenários, a segmentação dos grupos em indivíduos é difícil.

A principal contribuição deste trabalho é o desenvolvimento de novos algoritmos baseados em restrição homográfica que não necessitam de segmentação prévia dos indivíduos na imagem obtida por cada câmera. Ao invés de um método do tipo *segmentar-então-localizar*, propõe-se aqui uma abordagem do tipo *localizar-e-segmentar*, que integra a informação disponível de todas as câmeras antes de tomar qualquer decisão quanto à posição dos indivíduos em cena ou sua segmentação no plano de imagem de qualquer câmera.

1.2 Aplicações

A análise de cenas apresentando várias pessoas tem como função produzir sistemas capazes de obter automaticamente dados detalhados sobre movimentação, comportamento e intenção de indivíduos ou grupos. Diversas aplicações podem ser derivadas dessas informações.

Sistemas de vigilância

Uma aplicação natural encontra-se em sistemas de vigilância, um dos principais campos de pesquisa em visão computacional. Os algoritmos apresentados neste trabalho podem ser aplicados:

- na detecção e rastreamento de invasores de um determinado perímetro de segurança;
- na estimação da densidade de usuários em ambientes como estações de metrô e áreas comerciais, bem como seus padrões de locomoção (Figura 1.2.1);
- como etapa preliminar em sistemas de detecção de eventos e padrões de comportamento.

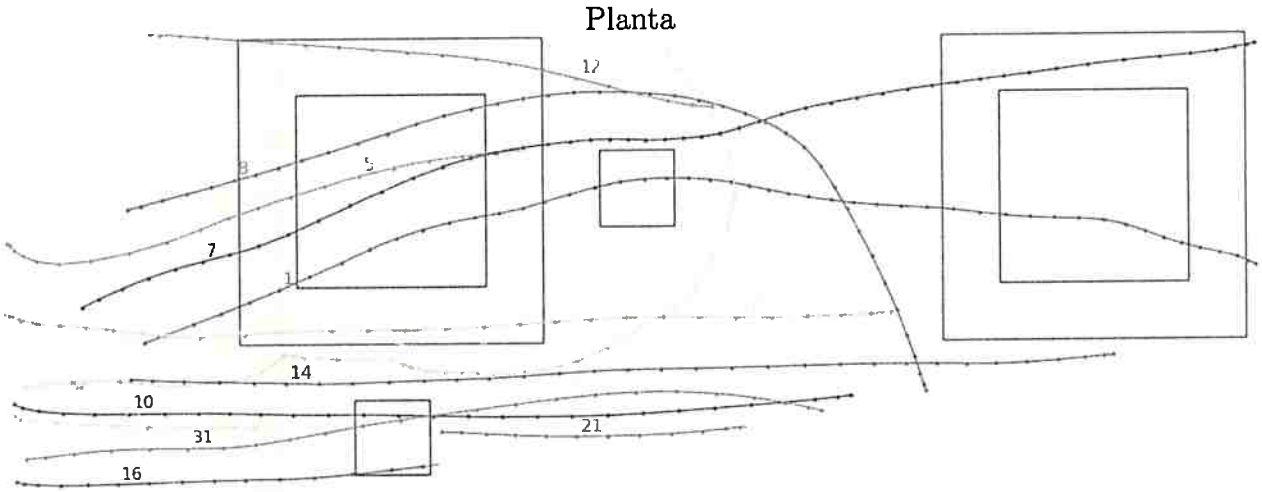


Figura 1.2.1: **Trajetórias individuais de pessoas vistas em uma estação de trem.** Trajetórias observadas em uma estação de trem (Victoria Station [48]), obtidas em um certo intervalo de tempo. A análise de resultados como este permite obter informações como áreas de aglomeração e trajetórias preferenciais. Mesmo alguns eventos simples poderiam ser identificados, como um grupo que se separa (indivíduos 1, 7 e 8) ou uma pessoa apresentando comportamento singular (indivíduo 2).

A Figura 1.2.1 exibe trajetórias de indivíduos vistos em uma estação de trem.

Ambientes atenciosos

Ambientes atenciosos são capazes de observar as atividades de seus usuários e prover serviços adequados. Detecção e rastreamento formam um componente perceptual desses sistemas, obtendo entrada dos sensores e fornecendo dados para modelos de situação [12]. Tem-se assim uma arquitetura que integra componentes de baixo nível (câmeras e outros sensores), médio nível (detecção e rastreamento) e alto nível (modelos de situação, eventos e atividades). O ambiente pode prover serviços baseados em quaisquer desses componentes, por exemplo, um controle de portas baseado exclusivamente nos sensores. Porém, serviços mais sofisticados dependem dos componentes de médio e alto nível, capazes de determinar estados mais complexos, resultantes da interação dos indivíduos entre si e com o ambiente.

3D-DTV e câmeras virtuais

Uma característica interessante aos futuros sistemas digitais de TV é o uso de 3 dimensões (3D-DTV). Tais sistemas poderiam permitir ao usuário a exploração da cena a partir de qualquer ponto de vista desejado, ou seja, vídeo com ângulo livre (*free view video*).

A localização e o rastreamento de pessoas permitem que esses objetos sejam removidos da cena. Um modelo 3D do ambiente, com informação de fundo (*background*), pode ser então construído e atualizado. Os indivíduos podem então ser reintroduzidos no ambiente, como ilustrado na Figura 1.2.2, ou mesmo substituídos por elementos sintéticos, como avatares.

1.3 Sumário

1.3.1 Contribuições

Este trabalho contribui com o estado da arte em detecção e rastreamento com múltiplas câmeras apresentando:

- uma nova técnica, baseada na transformada de Hough, desenvolvida para a localização de indivíduos em cena;
- um algoritmo baseado em múltiplos filtros de Kalman para rastreamento de múltiplos objetos, integrando as localizações obtidas em trajetórias individuais;
- o emprego do método proposto como estimador da altura de cada indivíduo - aborda-se também a aplicação dessa estimação na avaliação das relações de oclusão entre indivíduos e no rastreamento;
- vários testes com bases de dados públicas (PETS 2006 e PETS 2009) avaliados contra anotação manual, que foi produzida e disponibilizada.

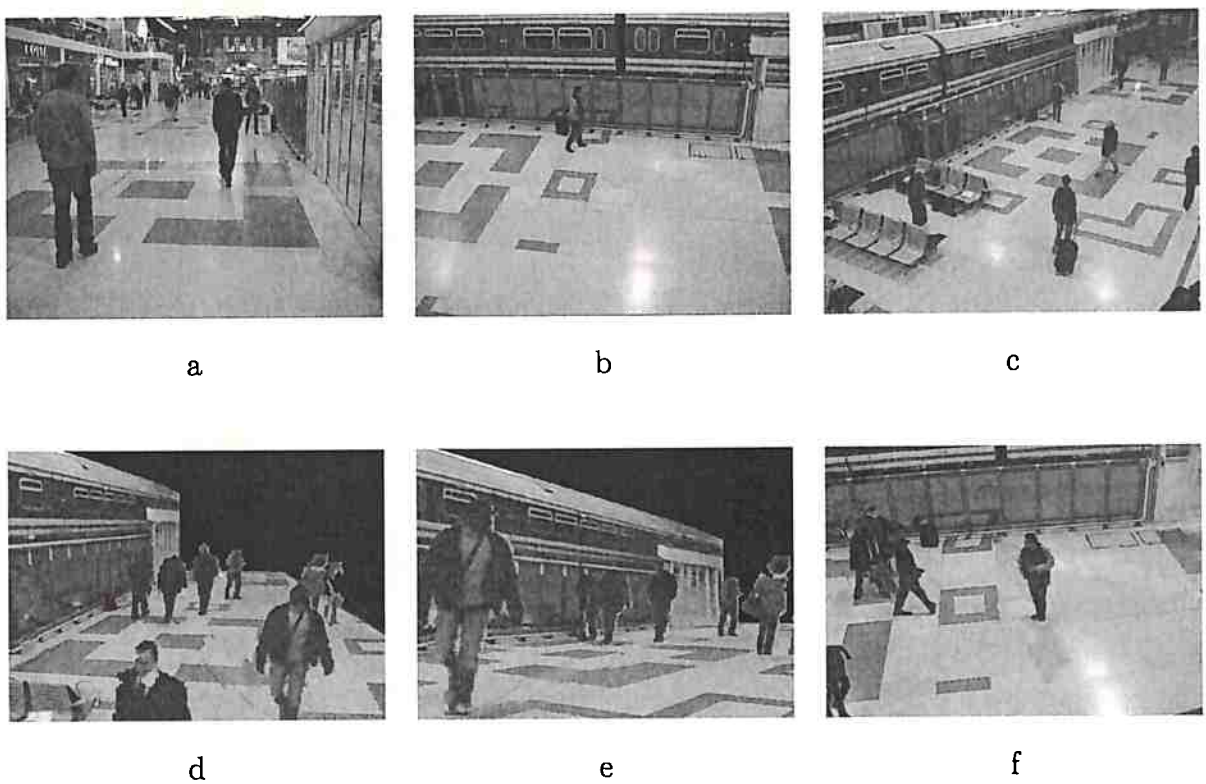


Figura 1.2.2: **Exemplo de vídeo com ângulo livre (*free view video*).** As imagens (a), (b) e (c) são provenientes de 3 câmeras. Já as imagens sintéticas (d), (e) e (f) exibem ângulos diferentes dos utilizados pelas câmeras reais. As pessoas foram removidas da cena através dos métodos apresentados nesta tese. Um modelo 3D do ambiente é gerado e texturizado utilizando-se o plano de fundo. Em seguida, os indivíduos são reintroduzidos, escolhendo-se a imagem real que apresenta ângulo mais próximo ao ângulo da câmera virtual. Figuras (d), (e) e (f) cortesia de Jeferson R. Silva.

1.3.2 Estrutura do texto

O texto prossegue da seguinte forma. O Capítulo 2 apresenta uma revisão dos trabalhos recentes nas áreas de subtração de fundo (*background subtraction*), detecção e rastreamento de múltiplos objetos utilizando uma ou mais câmeras.

No Capítulo 3, são introduzidas as notações empregadas ao longo do trabalho. São apresentadas as propriedades geométricas utilizadas e formas de estimar os elementos necessários, como matrizes de homografia e pontos afins. Assim como o método de subtração de fundo empregado é descrito e discutido.

O Capítulo 4 define o conceito de *suporte*, uma noção de "massa" dos objetos, oriunda da informação de *foreground* corrigida quanto a efeitos de perspectiva. É apresentado um algoritmo, baseado na Transformada de Hough, capaz de integrar o suporte encontrado em diversas câmeras diretamente em um planta do local imageado. Como é comum no arcabouço de Hough, os objetos de interesse correspondem a máximos locais encontrados no espaço paramétrico. O capítulo apresenta um método para a identificação desses máximos bem como um filtro para a remoção de falsos-positivos.

O Capítulo 5 descreve um método baseado no uso de múltiplos filtros de Kalman para rastrear os diversos indivíduos em cena, integrando no tempo os resultados da localização instantânea apresentados no capítulo anterior.

O Capítulo 6 exibe resultados obtidos na localização e rastreamento de pessoas para diversas bases de dados públicas, em situações de oclusão mútua.

O Capítulo 7 conclui a tese. São discutidas as vantagens e limitações do método proposto, caminhos para sua extensão e trabalhos futuros que podem ser derivados.

Capítulo 2

Trabalho relacionado

> *[...] descobertas científicas deveriam ser, de algum modo, inevitáveis. Elas devem estar no ar, produtos do clima intelectual de um tempo e lugar específicos. [...]*
>
> *Malcolm Gladwell* - In the air, *The New Yorker, Maio de 2008*

Rastreamento pode ser definido como o problema de estimar a trajetória de um objeto de interesse enquanto este se desloca pela cena, obtendo-se assim uma rotulação consistente dos objetos ao longo do tempo [55]. Dependendo da aplicação pretendida ou da solução adotada, a *predição* da localização de um objeto em um instante futuro pode ser também um requisito.

O rastreamento de pessoas em ambientes diversos, de amplos locais públicos a pequenas salas privadas, é de particular interesse. Trata-se de um pré-requisito para a maioria das aplicações que realizam *detecção de eventos*, tema sobre o qual a comunidade de visão computacional tem trabalhado intensamente. Esforços como as conferências *PETS* (*International Workshop on Performance Evaluation of Tracking and Surveillance*) [48, 18] e como a nova trilha de detecção de eventos no TRECVid 2008 [39] são alguns exemplos que ilustram a atenção que o tema tem recebido nos últimos anos.

Sistemas de vigilância, ambientes atenciosos, monitoramento de tráfego e navegação de veículos autônomos são algumas das aplicações que necessitam reconhecer eventos e realizar rastreamento robusto de objetos. Dentro das atividade de pesquisado do Laboratório de

Tecnologias para Interação (LaTIn) do IME-USP, o objetivo final é desenvolvimento de ambientes inteligentes com o uso de visão computacional. James Crowley [12] ilustra a questão:

> As tecnologias de informação e comunicação são autistas. Elas não apresentam qualquer noção sobre os papéis desempenhados pelos indivíduos durante suas interações sociais e não possuem habilidade em prever quais ações são apropriadas ou inapropriadas, nem qualquer senso sobre a perturbação que pode ser causada por um comportamento inapropriado durante a prestação de um serviço.

Reverter essa situação depende da identificação de eventos durante as interações dos usuários com as máquinas e entre si. Detecção e rastreamento robustos de objetos e indivíduos são um pré-requisito para tanto.

Um problema de grande importância é o tratamento de oclusões durante o rastreamento, seja ela causada por elementos da cena ou pelos próprios objetos entre si. No rastreamento de indivíduos, esta questão é marcante, pois a oclusão entre pessoas é frequente em qualquer agrupamento humano. Sejam colegas em uma sala de reuniões, sejam passageiros em uma estação de trem, o rastreamento robusto dos indivíduos depende de mecanismos próprios para o tratamento das oclusões que certamente irão ocorrem nessas situações.

Yilmaz *et al.* [55] apresentaram uma extensa revisão da literatura sobre rastreamento, abordando trabalhos até o ano de 2004 e fornecendo um largo panorama da área. O foco da presente tese é a rastreamento de objetos em oclusão mútua, particularmente humanos – com enfoque na componente de *detecção* dos objetos de interesse. Este capítulo irá apresentar uma revisão de alguns trabalhos da literatura recente que focam no tratamento de oclusão durante rastreamento de múltiplas pessoas. A Seção 2.1 apresenta em linhas gerais o arcabouço utilizado pela maioria dos métodos. A Seção 2.2 apresenta métodos que tentam tratar situações de oclusão empregando apenas uma câmera enquanto que abordagens que empregam diversos sensores são discutidos na Seção 2.3.

2.1 Ferramentas empregadas

A maioria dos métodos que serão apresentados nesse capítulo fazem uso de um esquema similar de rastreamento. Tal esquema consiste em:

1. definir os pixeis da imagem de entrada que pertencem aos objetos de interesse, separando-os dos pixeis que fazem parte do "fundo" da cena;
2. utilizar os pixeis selecionados para detectar objetos e
3. analisar os objetos detectados frente aos objetos presentes no histórico do rastreamento, definindo correspondências entre os objetos observados e os rastreados e atualizando as trajetórias obtidas.

A primeira etapa consiste em obter uma classificação para cada pixel das imagens de entrada, rotulando-o como *fundo* (*background*) ou *figura* (*foreground*), um ponto pertencente a um dos objetos de interesse. Esta classificação é conhecida na literatura como *subtração de fundo* (*background subtraction*) e será discutida em mais detalhes no Capítulo 3.

A terceira etapa pode ser realizada de diversas formas, incluindo correlação, heurísticas de mínima distância ou casamento entre grafos. Porém, os métodos largamente mais empregados são os métodos estatísticos de correspondência. Esses métodos empregam um modelo de *espaço de estados*, tradicionalmente utilizados na literatura de Teoria de Controle. O estado de um objeto contém propriedades como posição, velocidade e aceleração (em alguns modelos, atributos de forma são incluídos). São definidos também modelos para a *dinâmica* (a movimentação esperada de um objeto de um instante para outro) e para a *observação*, que relaciona estados às características observadas na imagem.

Os métodos mais utilizados na literatura para este fim são o *Filtro de Kalman* [36, 52] e, mais recentemente, os *Filtros de Partícula* (*particle filters*) [34, 29, 13]. O nome *filtro* vem de sua habilidade em reduzir ("filtrar") o espaço de busca através das distribuições de probabilidade impostas aos estados. Ambos os filtros apresentam etapas de *predição*, na qual o estado futuro do objeto é previsto com base no histórico corrente, aplicando-se o

modelo de dinâmica ao estado atual do objeto, e de *atualização*, onde o estado do objeto é atualizado a luz da informação observada, através do modelo de observação. No rastreamento de múltiplos objetos, uma única observação, oriunda da fase de detecção, precisa ser atribuída a cada filtro. Como resolver esse problema de associação é geralmente o foco da maioria dos métodos de rastreamento.

2.2 Soluções baseadas em um única câmera

Haritaoglu *et al.* [23] construíram um sistema, batizado de W^4, que opera em imagens em níveis de cinza obtidas por uma única câmera, capaz de rastrear múltiplos indivíduos em uma cena e tratar casos de oclusão. Eles empregam um modelo de distribuição bi-modal para a intensidade observada nos pixeis de fundo e formam componentes conexas (*blobs*) com os pixeis classificados como figura. Se o *blob* for classificado como um grupo em oclusão, o sistema irá segmentá-lo em indivíduos. A segmentação é realizada analisando pontos de grande curvatura nos *blobs* e o histograma produzido pela projeção dos pixeis na direção vertical, como ilustrado na Figura 2.2.1. Se estiverem distribuídos lado a lado, os indivíduos deveriam produzir um histograma onde os picos correspondem à localização de cada pessoa. O método contudo não é capaz de solucionar outras relações de oclusão, nas quais as pessoas estejam alinhadas verticalmente.

Para rastrear indivíduos, mesmo que eles estejam em grupos, W^4 emprega *modelos de aparência* (*appearance models*). Cada modelo de aparência é composto por um modelo de intensidade e uma máscara de probabilidade. O modelo de intensidade é uma região retangular exibindo a luminosidade observada para cada pixel do objeto. Já a máscara de probabilidade é definida sobre a mesma região, armazenando para cada pixel a probabilidade do objeto ser observado naquele pixel em particular. O rastreamento é realizado pela correlação entre os modelos observados e os armazenados para cada objeto sendo seguido. O sistema pode rodar em tempo real para entradas com resolução de 320×240, porém seu módulo de subtração de fundo é incapaz de lidar com sombras e outras variações de luminosidade.

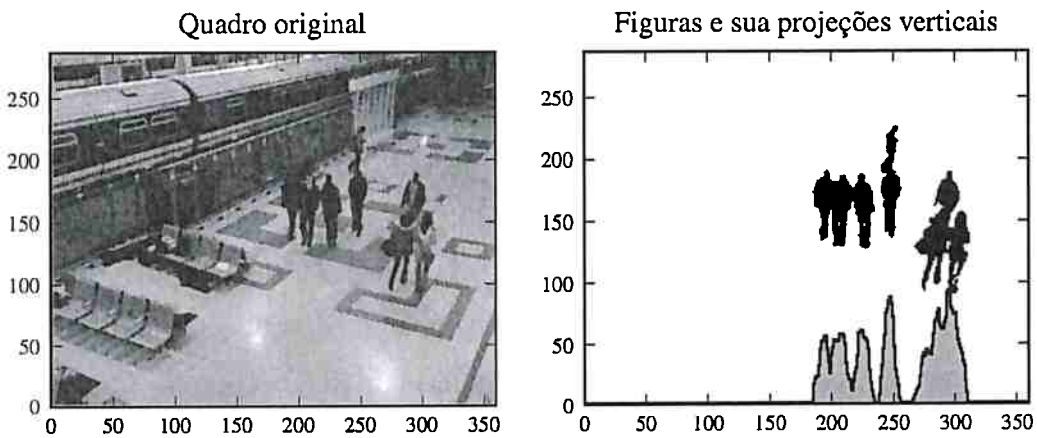


Figura 2.2.1: **Projeção das figuras na direção vertical**. Esta técnica, proposta por Haritaoglu *et al.* [23] em seu sistema W^4, visa separar indivíduos em oclusão. Os pixeis de figura são projetados verticalmente, formando um histograma. Assumindo que os indivíduos estejam em posição ereta, modas são esperadas próximas às regiões correspondentes a cabeça. No agrupamento da esquerda, os picos podem ser utilizados para separar os 3 indivíduos com sucesso. Contudo, outras relações de oclusão não são devidamente tratadas por este método, como pode ser visto nos dois agrupamentos restantes, onde os indivíduos em oclusão parcial não se encontram lado a lado.

Senior *et al.* [44] tratam situações de oclusão entre objetos identificando os instantes em que seus pixeis na imagem se unem em uma única componente conexa. O método também necessita de uma classificação figura/fundo de cada quadro de entrada. Os pixeis marcados como figura são agrupados em componentes conexas, que são ajustadas em janelas retangulares (*bounding boxes*). São computadas as distâncias entre as janelas observadas no instante corrente e as trilhas dos objetos sendo rastreados, formando-se uma matriz de associação. Se uma janela for associada a uma única trilha, é definida uma correspondência e a trilha é atualizada com a nova posição encontrada. Quando várias janelas são associadas a uma mesma trilha, uma oclusão é detectada. Similarmente, se uma única trilha for associada a várias janelas, então objetos que estavam em oclusão tornaram a se separar. Essas duas situações são tratadas com o auxílio de modelos de aparência.

O modelo de aparência proposto por Senior *et al.* é similar ao utilizado pelo W^4 de Haritaoglu *et al*, mas utiliza um modelo colorido em RGB. Quando uma trilha é criada, um novo modelo de aparência é inicializado. Esse modelo é atualizado ao longo do tempo conforme regiões vão sendo associadas à trilha. No momento em que uma oclusão ocorre, um classificador de máxima verossimilhança [15] é empregado para associar cada pixel a uma trilha, utilizando a informação dos modelos de aparência. Finalmente, avaliando-se a classificação obtida, relações de profundidade são definidas para os objetos, determinando-se assim a ordem de oclusão.

Isard e McCormick [30] optam por uma alternativa diferente do tradicional método sequencial definido por subtração de fundo, detecção e rastreamento. Eles propõem uma abordagem Bayesiana onde os três processos são realizados simultanemanete. Uma função de verossimilhança, derivada da teoria de correlação Bayesiana, é utilizada na comparação de hipóteses. Tais hipóteses se referem ao número de objetos presentes e suas configurações (forma e localização do objeto). A função é utilizada para definir um modelo de observação, onde se avalia a probabilidade de se obter a imagem observada dada um determinada configuração de objetos. Misturas de Gaussians são utilizados para obter modelos probabilísticos para a distribuições de cor do fundo e das figuras. Definido o modelo de observação,

o arcabouço de filtro de partículas é empregado. O filtro, através de sua etapa de predição, restringe o espaço de configurações a serem avaliadas pelo modelo de observação, sendo capaz de lidar com um número arbitrário e variável de objetos em cena e oclusões. As configurações com melhores estimativas são exibidas, compondo as trajetórias dos objetos. O método requer uma câmera calibrada e só é capaz de lidar efetivamente com oclusões se os modelos de figura puderem ser treinados.

Wu e Nevatia [54] optam por uma alternativa à classificação figura/fundo, baseando sua solução em características batizadas por eles de *edgelets*. *Edgelets* são segmentos curtos (entre 4 e 12 pixeis) representando linhas e curvas. Utilizando o arcabouço proposto por Viola e Jones [50], Wu e Nevatia desenvolvem, via AdaBoost [43], detectores para várias partes do corpo humano: (i) cabeça e ombros, (ii) torso e (iii) pernas. O problema de detecção de vários indivíduos é então formulado como um problema de máximo *a posteriori* que varre os espaço de soluções buscando pela configuração que fornece a melhor interpretação para as partes observadas. O rastreamento é realizado por associação, casando a melhor hipótese de localização com as detecções observadas. O método proposto dispensa calibração da câmera, mas o tamanho da janela de varredura para a detecção de partes não leva em consideração variações de escala ou efeitos de perspectiva.

2.3 Soluções baseadas em várias câmeras

Mittal e Davis [38] apresentam um sistema, batizado por eles de *M_2 Tracker*, que realiza detecção e rastreamento de pessoas utilizando múltiplas câmeras calibradas. O M_2Tracker representa cada indivíduo por sua posição no plano do solo e um modelo de aparência. Esse modelo de aparência inclui uma mapa de ocupação 2D, similar ao utilizado por Haritaoglu *et al.* [23], e um modelo de cor. O modelo de cor é formado por uma pilha de "fatias", faixas de igual largura definidas por sua distância (altura) ao solo. Cada fatia possui sua própria distribuição de probabilidade no espaço de cores, obtida por métodos não-paramétricos de estimação. Estes modelos são utilizados para segmentar as imagens provenientes de cada

câmera através de um classificador Bayesiano. Este classificador avalia a probabilidade de um certo pixel pertencer a cada objeto (ou ao fundo da cena), através das distribuições de cor de cada objeto e de um *prior* que avalia oclusão através dos mapas de ocupação. Uma vez segmentadas, as imagens das várias câmeras são combinadas duas as duas, varrendo-se as linhas epipolares para estabelecer assim um casamento entre regiões vistas nas duas imagens. Em outras palavras, identificam-se os segmentos, nas linhas epipolares, que correspondem a um mesmo objeto em duas imagens diferentes. Esses segmentos são utilizados para identificar pontos em 3D que residam no interior dos objetos, que são então projetados no plano do solo. Obtidas as projeções no solo, métodos não-paramétricos são utilizados novamente para estimar a localização do objeto para cada par de câmeras. As estimações são então combinadas através de uma análise de oclusão que permite atribuir pesos maiores a pares de câmeras que tenham uma visão livre de oclusões do sítio em questão. O procedimento é iterado, utilizando-se as posições encontradas, repetindo-se o processo até que as posições no solo se tornem estáveis. Finalmente, as posições encontradas são utilizadas para atualizar a posição dos indivíduos, através de Filtros de Kalman, e os modelos de aparência são então atualizados.

Fleuret *et al.* [19] empregam um arcabouço probabilístico para realizar simultaneamente detecção e rastreamento. O modelo empregado é uma combinação de um modelo de aparência com um modelo simples de movimento. O modelo de aparência é composto por uma densidade de probabilidade no espaço RGB e um mapa de ocupação no plano do solo. No mapa de ocupação, o plano do solo é particionado em uma grade e a probabilidade de ocupação de cada célula é estimada utilizando-se os resultados de um subtrator de fundo. Este modelo de ocupação é uma probabilidade condicional entre as figuras a configuração das células ocupadas. O Algoritmo de Viterbi é utilizado para obter a trajetória mais provável para cada indivíduo e uma heurística gulosa é aplicada de forma a otimizar uma trajetória por vez. Cada indivíduo precisa ser visto isoladamente ao menos um vez para a construção dos modelos. Calibração total das câmeras é também necessária.

2.3.1 Restrição homográfica

No ano de 2006 surgiram na literatura vários métodos capazes de integrar a informação oriunda de diversas câmeras sem necessitar de calibração total, isto é, sem a necessidade de conhecer todos o parâmetros, intrínsecos e extrínsecos de cada câmera. Esses métodos empregam apenas *homografias entre planos.* Uma homografia é uma transformação projetiva 2D que mapeia pontos de um plano para outro, preservando colinearidade (uma reta é mapeada para outra reta). Uma transformação homográfica possui 8 graus de liberdade e é inversível enquanto que a transformação de um câmera projetiva, de 3D para 2D, possui 11 graus de liberdade e não é inversível. Logo, operar com homografias significa que menos informações sobre cada câmera são necessárias. Homografias serão apresentadas e discutidas em mais detalhes no Capítulo 3. Por hora, será introduzido apenas o conceito fundamental para a compreensão dos métodos descritos a seguir.

A Figura 2.3.1 exibe uma cena onde três objetos repousam sobre o plano de imagem e são observados por três câmeras. O plano do solo induz uma homografia sobre os planos de imagem das três câmeras [24]. Isto significa que é possível mapear pontos entre o plano do solo e os planos das câmeras (ou entre os planos da câmeras) através de uma transformação linear de coordenadas homogêneas (ver Capitúlo 3). Pontos que pertencem às bases dos objetos, suas regiões de contato com o solo, são mapeados de forma consistente para os planos de imagem, isto é, são levados para a imagem dos objetos em todas as câmeras – como pode ser observado para os pontos **X**. **Y** e **Z** na figura. Pontos que não pertencem aos objetos ou pontos acima do plano do solo não gozam desta propriedade. Tal situação é ilustrada pelo ponto **W**, que é ocluso por objetos nas câmeras 1 e 2 mas é observado diretamente pela câmera 3, onde não é mapeado para a região de nenhum objeto.

Esta *restrição homográfica* tem sido empregada na literatura recente. Khan e Shah [31, 32] classificam os pixels da imagem de cada câmera como figura ou fundo. Eles exploram o fato de que só pixeis que correspondem às localizações dos indivíduos no plano do solo (os "pés") irão ser transformados de forma consistente para regiões de figura em todas as câmeras. Os autores modelam a distribuição de probabilidade dos pixeis de fundo e computam a proba-

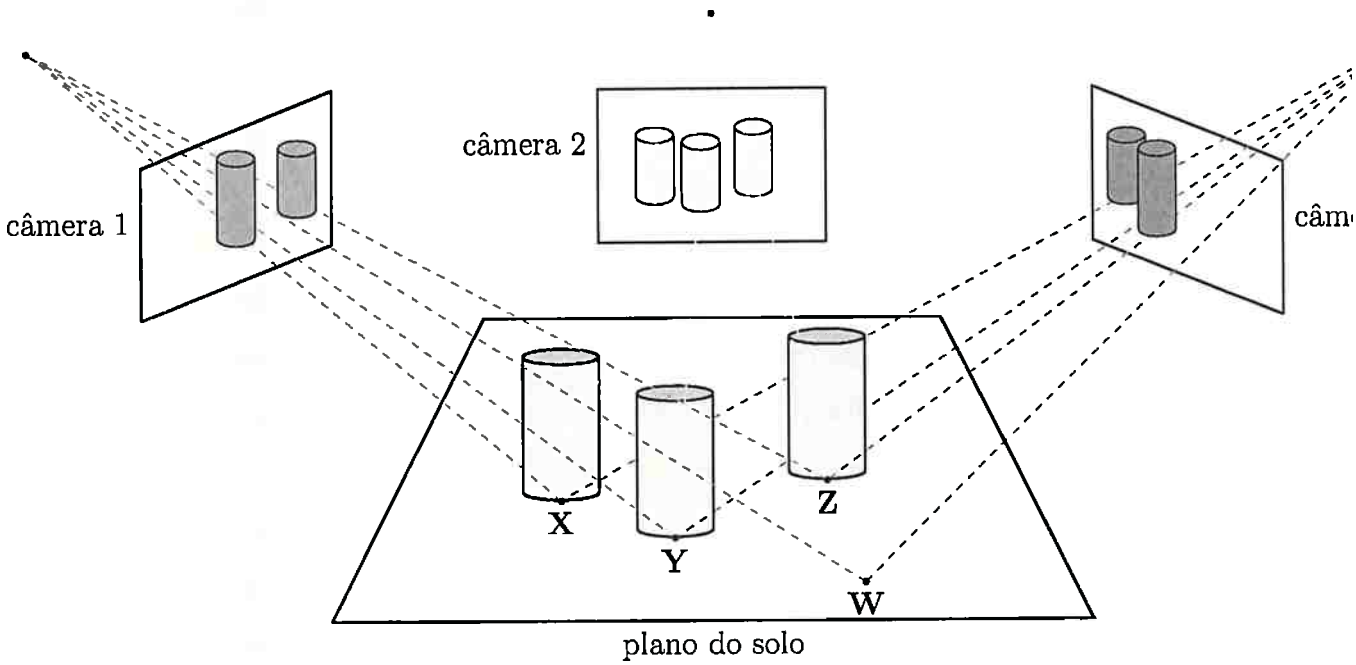


Figura 2.3.1: **Homografia induzida pelo plano do solo.** Três objetos são observados por três câmeras. Os pontos de contato dos objetos com o solo são sempre mapeados para pixeis classificados como figura no plano de imagem de cada câmera (assumindo que a subtração de fundo tenha sido realizada com sucesso), como pode ser observado para os pontos **X**, **Y** e **Z**. O mesmo não vale para pontos fora do plano do solo ou que não pertençam a nenhum objeto, como pode ser visto para o ponto **W**. Esse ponto está ocluso pelo objeto em **X** na câmera 1 e pelo objeto em **Z** na câmera 2, sendo então mapeado para pixeis de figura nessas duas câmeras. Porém, o mesmo não ocorre com a câmera 3, que discrimina a localização. Essa *restrição homográfica* vem sendo empregada frequentemente na literatura recente.

bilidade de um pixel pertencer a uma figura (objeto), produzindo um mapa de probabilidades para cada câmera. O plano de imagem de uma câmera é escolhido como plano de referência e as homografias induzidas pelo plano do solo são utilizadas para transformar os mapas de probabilidade de cada câmera para o mesmo plano de referência. Uma vez no mesmo espaço, os mapas de probabilidade são multiplicados para produzir o que os autores chamam de *mapa de sinergia*. Tal mapa é um produtório das probabilidades obtidas para cada câmera. Removem-se regiões que apresentam baixos valores pois, pela restrição homográfica, as localizações dos objetos corresponderão às regiões de alta sinergia. As imagens produzidas são então empilhadas ao longo do tempo, formando um volume 3D espaço $\times$ tempo. Este volume é segmentado pelo método de cortes normalizados [45]. Os autores assumem que as regiões correspondentes aos pés dos indivíduos são coerentes ao longo do tempo, de forma que os segmentos alongados produzidos pela segmentação corresponderão às trajetórias das pessoas em cena.

Hu *et al.* [27] assumem que os pixels que correspondem às figuras dos objetos são simetricamente distribuídos ao longo do eixo principal do mesmo. O eixo principal é utilizado por eles como a característica utilizada no rastreamento de indivíduos. Tomando-se as várias câmeras duas a duas, é possível obter um ponto de intersecção entre o eixo principal visto em uma câmera e a projeção do eixo principal observado na outra câmera, transformado através de uma homografia, obtendo-se assim um ponto no plano de imagem que corresponde a um ponto no plano solo, possível localização de um indivíduo. Para avaliar se os dois eixos correspondem ao mesmo indivíduo, os autores computam a distância entre a intersecção encontrada e a localização prevista por um Filtro de Kalman. Embora os autores façam uso da restrição homográfica, o método depende do rastreamento para resolver oclusões.

Kim e Davis [33] combinaram a idéia do cruzamento de eixos principais a um filtro de partículas para rastreamento de indivíduos. Primeiramente, um conjunto de partículas é obtido a partir do modelo de dinâmica. Em seguida, estes pontos são integrados aos modelos de aparência de cada indivíduo para produzir uma segmentação em cada plano de imagem. A segmentação e identificação dos indivíduos, a partir de seus modelos de cor e

das possíveis relações de oclusão, são similares ao procedimento utilizado no M_2Tracker [38]. Os eixos principais são computados e, uma vez que a identificação dos indivíduos define sua correspondência entre câmeras, as intersecções entre os eixos principais são obtidas, refinando a localização. O conjunto de partículas é então atualizado, de acordo com uma equação de observação. A desvantagem maior do método é que os indivíduos precisam ser observados inicialmente como objetos isolados, sem oclusão total ou parcial, de modo que seus modelos de aparência possam ser inicializados.

Eshel and Moses [17] utilizam a restrição homográfica induzida por diversos planos paralelos ao plano do solo, buscando pela localização das cabeças dos indivíduos em planos mais elevados. As imagens obtidas por cada câmera são transformadas para um plano de referência utilizando-se as homografias. Correlação entre a luminosidade dos pixeis é aplicada para identificar candidatos as cabeça. Em uma etapa preliminar de rastreamento, um método de associação baseado no vizinho mais próximo é utilizado para obter correspondências entre as regiões de cabeça ao longo do tempo. As trilhas encontradas são combinadas em trajetórias individuais através de uma heurística que combina seis medidas diferentes para avaliação de sobreposição, distância e direção entre trilhas. De acordo com os autores, pessoas vestidas com roupas de cores similares são uma fonte considerável de falsos-positivos, uma desvantagem natural no uso da correlação. As câmeras são colocadas em posições elevadas de modo a obter ângulos adequados, minimizando o efeito de oclusão. Os autores informam que o desempenho do sistema deteriora consideravelmente se menos do que cinco câmeras forem utilizadas.

A principal contribuição do trabalho apresentado nesta tese é um novo algoritmo, baseado na restrição homográfica, que não necessita de segmentação prévia dos objetos [23, 33, 27], nem empregar necessariamente a informação oriunda da predição obtida pelo rastreador [27]. Ao invés de uma abordagem do tipo *segmentar-então-localizar*, será proposto um método *localizar-então-segmentar*, integrando a informação de todas as câmeras disponíveis antes que qualquer decisão quando à localização de um objeto seja tomada. Uma visão geral do método será apresentada no próximo capítulo.

Capítulo 3

Descrição geral do sistema

> *[...] Dizendo background toda a gente sabe do que trata, mas não nos faltariam dúvidas se, em vez de background, tivéssemos chochamente dito plano de fundo, esse aborrecível arcaísmo, ainda por cima pouco fiel à verdade, dado que background não é apenas plano de fundo, é toda a inumerável quantidade de planos que obviamente existem entre o sujeito observado e a linha do horizonte. [...]*
>
> *José Saramago* - As intermitências da morte

O uso de múltiplos sensores na localização de diversos objetos em oclusão tem sido uma solução frequentemente empregada por métodos recentes, como visto no Capítulo 2. Uma das questões principais abordadas por esses métodos é o mecanismo de integração dos múltiplos sensores.

Tal integração deveria permitir que as várias câmeras compensassem mutuamente suas deficiências, de modo que oclusões insolúveis por uma câmera fossem resolvidas pelas demais. Idealmente, o método deveria ser computacionalmente simples, escalável para quantidades arbitrárias de câmeras e não impor restrições sobre a localização das mesmas.

O sistema proposto se divide em dois módulos: *detecção* e *rastreamento* (Figura 3.0.1). O módulo de detecção localiza indivíduos em uma planta da região observada, integrando a informação proveniente de vários sensores. O requisito mínimo é que cada região da planta

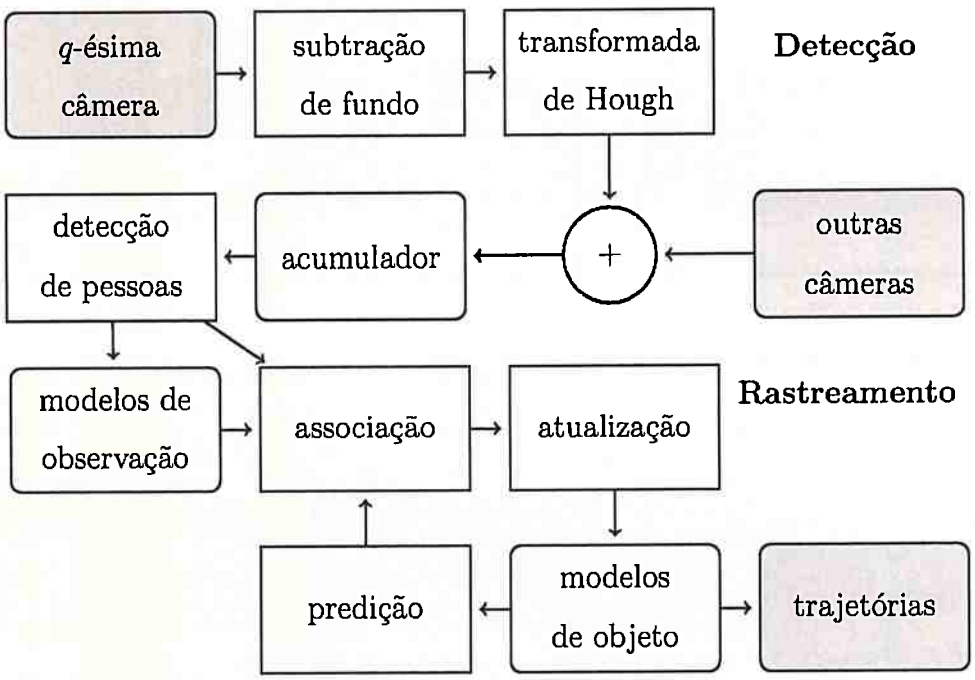


Figura 3.0.1: Diagrama de blocos do sistema proposto.

seja observada por ao menos dois sensores. As localizações encontradas servem como entrada para um módulo de rastreamento capaz de lidar com múltiplos objetos. Esse módulo usa o arcabouço clássico de predição-associação-atualização. As localizações encontradas pelo módulo anterior são utilizadas como observações, que são associadas às localizações previstas para os objetos em cena. Essas predições são então atualizadas, utilizando-se a observação associada. A evolução do estado do objeto produzem trajetórias, as posições ocupadas por ele ao longo do tempo.

O método proposto necessita que os pixeis da imagem obtida por cada sensor sejam previamente classificados como figura (*foreground*) ou plano de fundo (*background*). Neste trabalho, os sensores empregados foram câmeras coloridas e a classificação foi obtida por um algoritmo de subtração de fundo (*background subtraction*). Outras opções seriam câmeras infra-vermelhas, sonares ou imagens 2.5D.

A integração da informação proveniente das múltiplas câmeras é obtida através da Trans-

formada Generalizada de Hough. Um mapeamento é definido entre os pixeis que foram classificados como figura e as localizações na planta. Cada pixel é uma evidência que aponta para um conjunto limitado de localizações. O procedimento é repetido para cada câmera, integrando evidências sobre o mesmo acumulador. A restrição homográfica vista em vários trabalhos da literatura (Capítulo 2) garante que as localizações reais dos objetos acumulem evidência de forma consistente para todas as câmeras onde o objeto é visível. Finalmente, a detecção de pessoas é realizada buscando-se por localizações com grande acúmulo de evidência.

A Seção 3.1 discute o problema da subtração de fundo e apresenta o classificador adotado. A Seção 3.2 apresenta uma breve revisão dos fundamentos geométricos utilizados: representação de pontos em coordenadas homogêneas, transformações projetivas 2D (homografias) e pontos ideais. A integração da informação oriunda de múltiplas câmeras é realizada através da restrição homográfica, discutida na Seção 6.2. Métodos para estimação das matrizes e homografia são discutidos na Seção 3.4. Finalmente, a Transformada de Hough é introduzida brevemente na Seção 3.5. Ela será discutida novamente no Capítulo 4, quando suas propriedades serão exploradas na detecção de pessoas, integrando a informação obtida pela subtração da fundo através de relações geométricas apresentadas neste capítulo.

3.1 Subtração de fundo

O métodos desenvolvido utilizou como entrada imagens provenientes de várias câmeras coloridas fixas. Porém, a teoria apresentada pode ser aplicada a qualquer dispositivo que realize uma transformação projetiva que transforme a informação 3D do ambiente para dados em um plano 2D. Além da natureza projetiva do sensoreamento, é necessária uma classificação prévia dos pixeis como *figura* ou *fundo* (*foreground*/*background*).

Tal classificação visa apenas definir quais pixeis na imagem pertencem a objetos de interesse. Embora os conceitos de figura e fundo dependam fortemente da aplicação em questão, o problema foi generalizado e abordado na literatura como *subtração de fundo* (*background*

subtraction), concentrando-se sobretudo na identificação de objetos em movimento em aplicações de vigilância [47, 49, 5].

Subtração de fundo é um processo que consiste de duas etapas: a construção de um *modelo de fundo* e a classificação dos pixeis observados em um certo instante. A maioria dos sistemas adota um modelo para cada pixel da imagem.

Toyama *et al.* [49] analisaram o problema, listando algumas situações comuns a vários dos sistemas de subtração de fundo, que podem se tornar fontes de *falsos-positivos* (fundo classificado como figura) e/ou *falsos-negativos* (figura classificada como fundo):

objeto movido se um objeto que parte do fundo for movido, ele não deveria ser classificado como figura – ao menos não por tempo indeterminado;

"hora do dia" mudanças graduais de iluminação decorrentes da luz ambiente não deveriam produzir falsos-positivos;

acender/apagar de luzes (*light switch*) mudanças bruscas de iluminação, comuns em iluminação de interiores, afetam drasticamente a imagem produzindo falsos-positivos;

"árvores ao vento" (*waving trees*) [1] o fundo pode apresentar alguma dinâmica, como arbustos se movendo ao vento ou ondas em um espelho d'água;

camuflagem os pixeis do objeto são similares aos pixeis de fundo, produzindo falsos-negativos;

treinamento (*bootstrapping*) dados para o aprendizado do modelo de fundo podem não ser disponíveis;

abertura quando um objeto de cor homogênea se move, seus pixeis internos podem não ser detectados;

sombras os objetos podem projetar sombra sobre o fundo, fazendo com que ele difira do modelo e produza falsos-positivos;

[1] Preferiu-se aqui manter a denominação original por ser frequentemente utilizada na literatura.

objeto imóvel (*sleeping person*) um objeto que se torna estático, um problema dual do objeto movido.

objeto desperto (*waking person*) um objeto outrora estático torna a se mover, causando erros de classificação tanto nos pixeis de figura como no de fundo.

Toyama *et al.* argumentam que, em geral, a subtração de fundo deveria ser um processo de classificação pixel-a-pixel. A alternativa, classificação baseada em regiões da imagem, implicaria na resolução de problemas de segmentação de objetos, que são mais difíceis – uma tentativa de obter informação semântica a partir de visão de baixo nível.

Várias formas de modelar a distribuição de cor de um pixel foram empregadas na literatura. Toyama *et al.* [49] utilizam um filtro de Wiener [53], obtendo um predição linear do valor do pixel a partir dos valores observados em quadros anteriores. Visando obter alto-desempenho através de poucas comparações e adições de números inteiros, Boult *et al.* [5] produziram um sistema que armazena apenas dois valores para cada pixel. A atualização dos valores é realizada através da adição (ou subtração) de uma constante inteira, caso o valor armazenado seja inferior (ou superior) ao valor observado para cada pixel. Os dois valores representam dois modelos de fundo, visando tratar bi-modalidade na distribuição.

Para tratar multi-modalidade, Stauffer e Grimson [47] modelam a distribuição de cada pixel através de uma mistura de Gaussianas. A forma tradicional de obter tal modelo é através do método de *maximização de esperança* (*expectation-maximization* – EM) [15, 6]. Porém, dado o alto custo computacional desse método de otimização, os autores optaram por uma método de atualização *on-line*. Em um certo instante t cada uma das K Gaussianas utilizadas é representada por um peso $\omega_{k,t}$, uma média $\mu_{k,t}$ e uma matriz diagonal de covariância $\sigma^2_{k,t}$. Se o valor v_t observado pertencer a um dos modelos Gaussianos (sua diferença absoluta à média é inferior a 2.5 vezes o desvio padrão), o pixel é classificado como fundo e o modelo Gaussiano é atualizado utilizando-se

$$\omega_{k,t+1} = (1 - \alpha)\omega_{k,t} + \alpha, \tag{3.1}$$

$$\mu_{t+1} = (1 - \rho)\mu_{k,t} + \rho v_t, \tag{3.2}$$

$$\sigma_{t+1}^2 = (1-\rho)\sigma_{k,t}^2 + \rho(v_t - \mu_{k,t})^\top (v_t - \mu_{k,t}), \tag{3.3}$$

onde α é uma *razão de aprendizado* e ρ é uma segunda razão derivada de α através de

$$\rho = \alpha \cdot N(I_t | \mu_{k,t}, \sigma_{k,t}), \tag{3.4}$$

sendo $N(\cdot)$ a distribuição normal.

3.1.1 Solução adotada

Os métodos que serão apresentados no Capítulo 4 podem fazer uso de qualquer método de subtração de fundo. Eles são particularmente robustos em relação a falsos-positivos obtidos pela classificação figura/fundo, mas são sensíveis a grandes quantidades de falsos-negativos, geralmente associados à camuflagem.

Misturas de Gaussianas foram adotadas para modelar cada pixel. Ao invés de adotar-se um procedimento *on-line* similar à atualização linear utilizada por Stauffer e Grimson [47], o método de EM foi empregado na estimação dos modelos Gaussianos. Esta escolha se justifica por dois motivos principais. Primeiro, há sequencias de treinamento para estimação de modelos de fundo disponíveis nas bases de dados utilizadas nos experimentos [48, 18]. Segundo, alguns dos indivíduos observados em cena apresentam um padrão errático de movimento, mantendo-se praticamente imóveis por intervalos arbitrários de tempo, o que torna mais difícil a escolha de uma razão de aprendizado adequada.

Seja uma imagem uma função $I : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N} \times \mathbb{N}$, que mapeia pontos (x, y) em um grade de pixeis em uma tupla RGB, cada componente (canal) sendo um inteiro pertencente ao intervalo $[0, 255]$ (8 bits). Assuma que os valores dos canais para um pixel $\mathbf{x} = (x, y)$ são representados por $I_R(\mathbf{x})$, $I_G(\mathbf{x})$ e $I_B(\mathbf{x})$.

Valores em RGB são sensíveis a pequenas mudanças de iluminação, como as causadas por sombras. Por esse motivo, é mais comum empregar componentes *normalizadas*, que apresentam menor variação:

$$I_r(\mathbf{x}) = \frac{I_R(\mathbf{x})}{I_R(\mathbf{x}) + I_G(\mathbf{x}) + I_B(\mathbf{x})}, \tag{3.5}$$

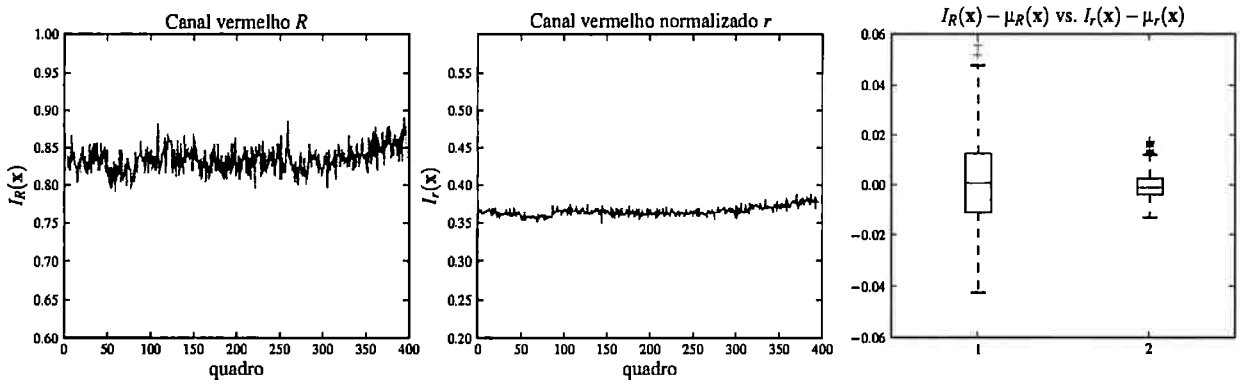


Figura 3.1.1: **Diferença de sensibilidade à luminosidade.** Os valores exibidos são os observados para o pixel $\mathbf{x} = (350, 190)$ em um conjunto de 400 imagens. Essas imagens correspondem à câmera 1 no conjunto de treinamento para subtração de fundo do PETS 2009 [18]. A variação observada no canal R é visivelmente maior que a observada para o canal normalizado r.

$$I_g(\mathbf{x}) = \frac{I_G(\mathbf{x})}{I_R(\mathbf{x}) + I_G(\mathbf{x}) + I_B(\mathbf{x})}, \tag{3.6}$$

$$I_b(\mathbf{x}) = \frac{I_B(\mathbf{x})}{I_R(\mathbf{x}) + I_G(\mathbf{x}) + I_B(\mathbf{x})}. \tag{3.7}$$

A Figura 3.1.1 exibe a diferença na variação dos valores observados para R e r em uma sequência de treinamento do PETS 2009, deixando clara as vantagens no uso das componentes normalizadas. Contudo, o uso das três componentes eliminaria a informação de intensidade luminosa, indispensável para a caracterização da imagem. Assim, adota-se o espaço rgL onde L representa a luminosidade:

$$I_L(\mathbf{x}) = \frac{\min[I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})] + \max[I_R(\mathbf{x}), I_G(\mathbf{x}), I_B(\mathbf{x})]}{2}. \tag{3.8}$$

Dado um pixel $\mathbf{x}$ e um conjunto de N imagens de treinamento $\{I_i\}_{i=1..N}$, o algoritmo de maximização da esperança [6] pode ser utilizado para estimar os pesos $\omega_k(\mathbf{x})$, as médias $\mu_{k,c}(\mathbf{x})$ e a variância $\sigma^2_{k,c}(\mathbf{x})$ em cada canal $c = r, g, L$ para cada modelo Gaussiano $k = 1..K$. O número K de Gaussianas empregadas é escolhido de forma a cobrir a multi-modalidade observada na distribuição do pixel. A Figura 3.1.2 exibe o modelo encontrado para o fundo da câmera 1 na base de dados PETS 2009.

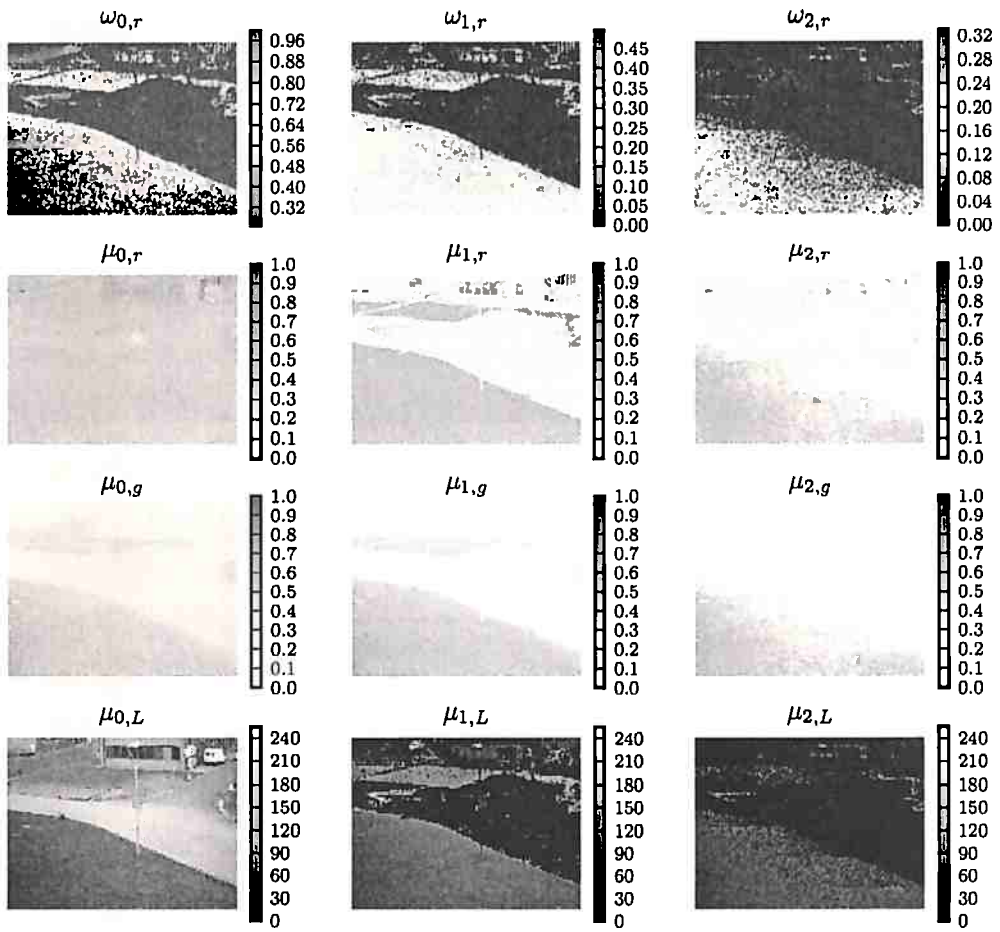


Figura 3.1.2: **Distribuição dos pixeis de fundo como uma mistura de Gaussianas.** A figura exibe o modelo com 3 Gaussianas para a câmera 1 do PETS 2009 [18], obtido através do algoritmo EM [6]. A primeira linha exibe os pesos de cada modelo. As linhas restantes exibem a média obtida para os canais r (vermelho normalizado), g (verde normalizado) e L (luminosidade). É possível notar que o asfalto visto em cena apresenta distribuição mono-modal enquanto o gramado exibe uma distribuição multi-modal.

Classificação

A classificação emprega dois parâmetros: $\omega_{\min}$ e α. Um pixel $\mathbf{x}$ em uma imagem I é classificado como fundo se existir uma Gaussiana k tal que $\omega_k(\mathbf{x}) \geq \omega_{\min}$ e

$$|I_c(\mathbf{x}) - \mu_{k,c}(\mathbf{x})| \leq \alpha \cdot \sigma_{k,c}(\mathbf{x}) \tag{3.9}$$

para todo canal $c = r, g, I$. Intuitivamente, um pixel é considerado fundo se ele puder ser atribuído a um modelo suficientemente representativo.

É razoável assumir que a luminosidade apresentada por um pixel de fundo possa decair, até um certo limite mínimo, devido a sombras projetadas por objetos em trânsito na cena ou por pequenas variações da iluminação ambiente. Assim, adota-se aqui a solução proposta por Wang e Suter [51]: se um pixel de fundo encontra-se sob sombra, espera-se que, para algum modelo k, o valor observado $I_L(\mathbf{x})$ esteja restrito a

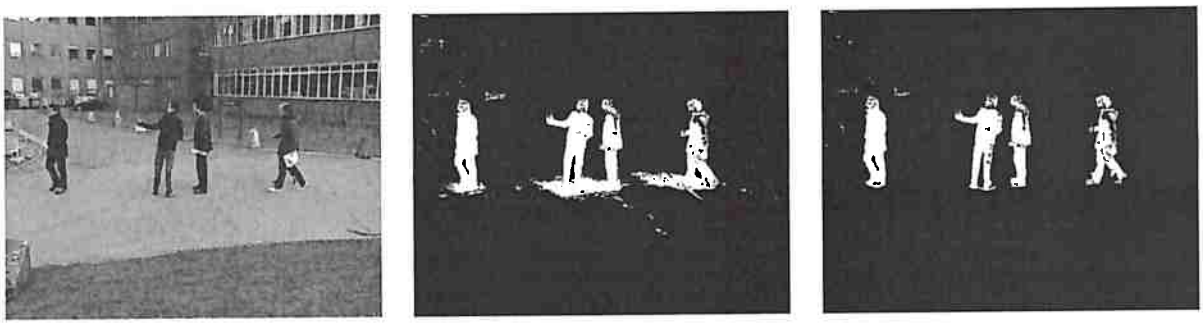
$$\beta \leq \frac{I_L(\mathbf{x})}{\mu_{k,L}(\mathbf{x})} \leq 1 \tag{3.10}$$

onde β é uma constante pré-definida. A classificação se dá então avaliando-se a Equação 3.9 para os três canais. Caso o pixel obedeça os modelos de cromaticidade (r e g) mas falhe em relação à luminosidade, o teste da Equação 3.10 é aplicado, decidindo-se assim se o pixel em questão deve ser classificado como figura ou se é apenas uma situação de sombra sobre fundo. A Figura 3.1.3 ilustra a diferença obtida ao se utilizar tal método.

3.2 Geometria

Para a compreensão dos métodos que serão apresentados no capítulos seguinte, é necessário estabelecer algumas bases sobre a geometria do imageamento que será empregada, bem como a notação utilizada.

O modelo de câmera utilizado é o *modelo de projeção central*, no qual pontos no espaço são projetados no plano de imagem através de raios que atravessam um mesmo ponto, o *centro de projeção*. Sejam $\mathbf{X}$ e $\mathbf{C}$ dois pontos no espaço. A projeção de $\mathbf{X}$ no plano de



a b c

Figura 3.1.3: **Remoção de sombras em subtração de fundo.** (a) Quadro original proveniente do PETS 2009 (câmera 5). (b) Resultado da classificação obtida aplicando-se a Equação 3.9 e modelo de mistura de Gaussianas. É possível ver grandes regiões de falsos-positivos junto aos pés das pessoas, causados pelas sombras que os mesmos projetam no solo. (c) resultado obtido aplicando-se o método de Wang e Suter [51] com $\beta = 0.7$. As sombras foram totalmente removidas.

imagem de um câmera com centro de projeção em $\mathbf{C}$ é o ponto de intersecção $\mathbf{x}$ entre o plano de imagem e a linha $\langle \mathbf{X}, \mathbf{C} \rangle$ definida por $\mathbf{X}$ e $\mathbf{C}$, como visto na Figura 3.2.1.

São introduzidas aqui duas notações que serão utilizadas ao longo de todo o texto. Pontos serão representados por letras em negrito, sendo as maiúsculas utilizadas para pontos quaisquer no espaço e minúsculas para pontos que se encontram sobre o plano de imagem (ver Tabela 3.2). Retas no plano de imagem também serão representadas por letras minúsculas em negrito.

3.2.1 Coordenadas homogêneas

O uso de coordenadas homogêneas é uma alternativa elegante para representação de pontos e apropriada ao estudo de transformações projetivas, uma vez que diversas propriedades geométricas podem ser obtidas através de operações simples da álgebra linear [24].

Considere uma reta em sua representação cartesiana $ax + by + c = 0$. Esta reta pode ser representada através de um vetor coluna $\mathbf{l} = (a, b, c)^\top$. Se um ponto (x, y) for representado por um vetor coluna da forma $\mathbf{x} = (x, y, 1)^\top$, é possível determinar se esse ponto pertence a

Q	Número de câmeras disponíveis
Π	Plano do solo, correspondente à planta do local analisado
$\mathbf{x}, \mathbf{y}$	Pontos em um plano de imagem qualquer, em coordenadas homogêneas – $\mathbf{x} = (x, y, 1)^\top$
$\mathbf{x}^q, \mathbf{y}^q$	Pontos no plano de imagem da q-ésima câmera, em coordenadas homogêneas
$\mathbf{l}$	Linha $ax + by + c = 0$, em coordenadas homogêneas – $\mathbf{l} = (a, b, c)^\top$
$\vec{\mathbf{n}}$	Vetores em coordenadas homogêneas – corresponde a um ponto ideal
$\mathbf{X}, \mathbf{Y}$	Pontos no plano do solo Π, em coordenadas homogêneas – $\mathbf{X} = (X, Y, 1)^\top$
$\mathbf{v}_Z^q$	Ponto afim em relação ao eixo Z (ortogonal ao plano Π)
$\langle \mathbf{x}, \mathbf{y} \rangle$	Linha entre os pontos $\mathbf{x}$ e $\mathbf{y}$. Pode ser obtida pelo produto cruzado $\mathbf{x} \times \mathbf{y}$
$\mathtt{H}$	Matrizes
$\mathbf{l}_\infty$	Linha no infinito ("horizonte")
$\mathbf{x}_\perp, \mathbf{x}_\top$	Convenções utilizadas para indicar pontos que correspondem à base ($\mathbf{x}_\perp$) e ao topo ($\mathbf{x}_\top$) de um objeto no plano de imagem, em relação ao vetor normal do plano Π

Tabela 3.1: **Notações utilizadas neste trabalho**. As notações adotadas aqui são inspiradas nas utilizadas por Hartley e Zisserman [24] e Criminisi [11], modificadas para se adequarem ao presente contexto.

l através do produto interno

$$\mathbf{x}^\top \mathbf{l} = 0, \tag{3.11}$$

que simplesmente remete à representação cartesiana da reta.

Para qualquer constante $k \neq 0$, as equações $ax + by + c = 0$ e $(ka)x + (ka)y + (kc) = 0$ definem a mesma reta, de forma que os vetores $(a, b, c)^\top$ e $k(a, b, c)^\top$ são equivalentes. O termo *vetor homogêneo* se refere a classe de equivalência representada por um vetor $(a, b, c)^\top$. Como $(kx, ky, k)^\top \mathbf{l} = 0$ se e somente se $(x, y, 1)^\top \mathbf{l} = 0$ é natural considerar que a classe de equivalência definida por $(kx, ky, k)^\top$ represente o ponto (x, y). Assim, um vetor homogêneo $\mathbf{x} = (x_1, x_2, x_3)^\top$ representa o ponto $(\frac{x_1}{x_3}, \frac{x_2}{x_3})$ em $\mathbb{R}^2$ [24].

A representação em coordenadas homogêneas de uma reta $\mathbf{l}$ definida por dois pontos $\mathbf{x}, \mathbf{y}$ pode ser obtida pelo produto cruzado

$$\mathbf{l} = \mathbf{x} \times \mathbf{y}. \tag{3.12}$$

Similarmente, o ponto de intersecção $\mathbf{x}$ entre duas retas $\mathbf{m}$ e l pode ser obtido por

$$\mathbf{x} = \mathbf{m} \times \mathrm{l}, \tag{3.13}$$

com a vantagem de que, se as retas forem paralelas, teremos um *ponto no infinito* na forma $(x, y, 0)^\top$. Embora o ponto $(x/0, y/0)$ não exista em $\mathbb{R}^2$, ele remete a idéia usual de que retas paralelas se encontram no infinito. Pontos da forma $(x_1, x_2, 0)^\top$ são também chamados de *pontos ideais*. Note que os conjunto dos pontos ideais forma um única linha, a *linha no infinito* $\mathrm{l}_\infty = (0, 0, 1)^\top$. O espaço formado pela união de $\mathbb{R}^2$ com o conjunto de pontos no infinito forma o *espaço projetivo* $\mathbb{P}^2$.

3.2.2 Transformações projetivas

Uma *transformação projetiva* é um mapeamento inversível de pontos em $\mathbb{P}^2$ para pontos em $\mathbb{P}^2$ que mapeia linhas em linhas, preservando assim colinearidade entre pontos. Uma definição algébrica para estsa transformação em $\mathbb{P}^2$ é dada pelo seguinte teorema [24]:

Teorema 3.2.1. *Um mapeamento $h : \mathbb{P}^2 \rightarrow \mathbb{P}^2$ é uma transformação projetiva se e somente se existe uma matriz 3×3 não singular $\mathtt{H}$ tal que, para todo ponto $\mathbf{x} \in \mathbb{P}^2$, $h(\mathbf{x}) = \mathtt{H}\mathbf{x}$.*

Hartley e Zisserman [24] apresentam parte da prova do Teorema 3.2.1. Um tratamento mais formal para transformações projetivas pode ser visto em Semple e Kneebone [21]. Essas transformações projetivas em $\mathbb{P}^2$ são também chamadas de *homografias*.

A projeção central pode ser vista como uma homografia entre planos no espaço. Particularmente útil é o mapeamento entre uma plano de referência qualquer no espaço e o plano de imagem, como ilustrado na Figura 3.2.1. Tal mapeamento pode ser representado como um mapeamento linear de coordenadas homogêneas, na forma de uma multiplicação por uma matriz $\mathtt{H}$.

O objetivo deste trabalho é a localização de objetos em cena. Assumindo que tais objetos repousem sobre uma superfície plana no solo, representada aqui pelo plano Π (Figura 3.2.1), existem então matrizes $\mathtt{H}$ que podem ser utilizadas no mapeamento da localizações dos objetos em Π em ponto do plano de imagem.

Seja XYZ uma base do espaço $\mathbb{R}^3$ tal que o plano Π seja dado dado por $X = 0$, $Y = 0$ e que Π contenha a origem do espaço O. Assuma que a direção de Z seja a mesma do vetor normal de Π. Logo, as localizações em $\mathbb{R}^3$ são pontos $\mathbf{X}$ da forma $(x_1, x_2, 0)$. Dado que a terceira componente do vetor é fixa, podemos representar os pontos de interesse na cena em $\mathbb{R}^3$ como pontos no plano Π com coordenadas em $\mathbb{P}^2$ e assim definir uma mapeamento linear de coordenadas homogêneas entre Π e o plano de imagem. Essa transformação será utilizada extensivamente no Capítulo 4.

3.2.3 Pontos ideais

Como todas as retas paralelas em uma direção qualquer se encontram no mesmo ponto no infinito, tal ponto torna-se naturalmente uma representação desta direção. Esses pontos ideais são então apropriados para representar direções no espaço.

A propriedade essencial das transformações projetivas é que elas são capazes de alterar a posição da linha no infinito de modo que um ponto ideal $\mathbf{x}_\infty$ no plano original Π pode ser

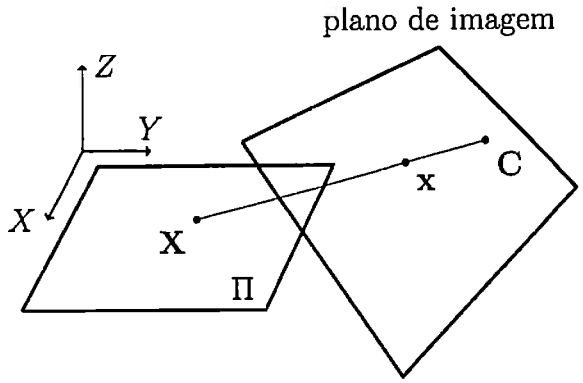


Figura 3.2.1: **Homografia entre um plano de referência o plano de imagem.** O plano de referência Π, que corresponde ao plano do solo, é imageado por um câmera com centro de projeção em $\mathbf{C}$. O mapeamento entre uma localização $\mathbf{X}$ no solo e sua imagem $\mathbf{x}$ pode ser realizado por uma transformação linear de coordenadas homogêneas $\mathbf{X} = \mathtt{H}\mathbf{x}$.

mapeado para um ponto $\mathbf{x}' = \mathtt{H}\mathbf{x}_\infty$ em Π' que *não é necessariamente um ponto ideal.* Isto significa que a informação sobre uma direção poderá ser obtida a partir de um ponto comum em $\mathbb{R}^2$ observado em Π'. O uso de tais pontos é bem conhecido por desenhistas: eles são os *pontos de fuga*, utilizados para guiar o traço do artista na representação de linhas paralelas de uma cena, como as paredes de um edifício ou as bordas de um caixote.

A importância dos pontos de fuga aqui é determinar a direção de retas observadas no plano de imagem. Mais exatamente, os dois elementos que serão utilizados neste trabalho são (i) o ponto $\mathbf{v}_Z$, o ponto afim da direção normal ao plano do solo Π e (ii) a *linha do horizonte* $\mathbf{l}_\infty$, a projeção da linha no infinito de Π no plano de imagem. O ponto $\mathbf{v}_Z$ é utilizado para determinar, para cada ponto no plano de imagem, qual a direção vertical, determinando assim a orientação esperada para o eixo principal de um objeto (uma pessoa em postura ereta, por exemplo). Já a linha do horizonte, em conjunto com o ponto $\mathbf{v}_Z$, será utilizada para compensar efeitos de perspectiva, como será discutido no Capítulo 4. A linha $\mathbf{l}_\infty$ impõe ainda uma restrição importanto quanto à localização de um objeto: no plano de imagem, nenhum ponto pertecente ao plano Π pode ser imageado acima da linha do horizonte. A Figura 3.2.2 ilustra as propriedades desses dois elementos. O ponto $\mathbf{v}_Z$ é a projeção do

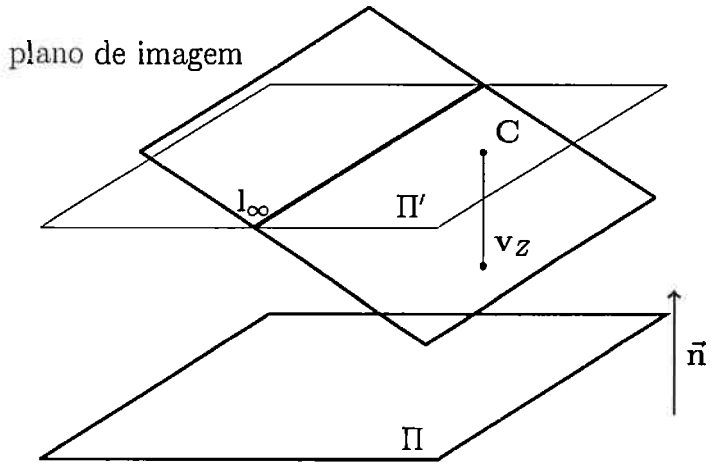


Figura 3.2.2: **Ponto afim da direção normal v_Z e a linha do horizonte l_∞.** O ponto afim v_Z é a projeção de **C** sobre o plano de imagem na direção normal de Π. A linha ideal l_∞ de Π foi transformada para uma linha comum no plano de imagem, dada pela intersecção deste com Π′, o plano paralelo a Π que contém **C**. Se Π corresponder ao plano do solo, então a reta l_∞ é a *linha do horizonte.*

próprio centro de projeção **C** sobre o plano de imagem, na direção do vetor normal do plano Π. Similarmente, a linha l_∞ é obtida pela interseção entre o plano de imagem e Π′, o plano paralelo a Π que contém o centro de projeção.

Pontos afins podem ser obtidos identificando-se linhas no plano de imagem que correspondam à direção de interesse. No caso de v_Z, identificam-se duas linhas "verticais", ou seja, ortogonais ao plano do solo Π. Embora sejam paralelas na cena, no plano de imagem elas podem surgir como retas concorrentes (Figura 3.2.3). Quando há mais retas disponíveis, métodos de estimação mais elaborados podem ser empregados [3, 37, 46, 24].

Existe porém um exceção, quando um ponto ideal é mapeado para outro ponto ideal no plano de imagem. No caso de v_Z, isto ocorre quando o plano de imagem é ortogonal ao plano do solo. Entretanto, não se trata de uma desvantagem: significa apenas que há um único vetor no plano de imagem que define a direção normal a Π para todos os pontos (na prática, se a câmera foi posicionada de maneira regular, significa que o eixo das ordenadas define a vertical da cena).

A linha do horizonte $\mathbf{l}_\infty$ pode ser obtida de maneira similar. Identificam-se dois pares de linhas no plano de imagem que correspondam a linhas "horizontais", paralelas ao plano do solo Π, mas que são imageados como pares de linhas concorrentes. Cada par de linhas fornece um ponto da linha do horizonte. Tomando-se os dois pontos, define-se a linha $\mathbf{l}_\infty$. Havendo múltiplos pares, estimadores mais elaborados podem ser empregados, como no caso de $\mathbf{v}_Z$. A Figura 3.2.3 ilustra a idéia.

3.3 Restrição homográfica

O plano do solo Π pode ser transformado para o plano de imagem e vice-versa graças às homografias. Mas o que ocorre com um ponto $\mathbf{X}'$ no espaço 3D que *não* pertence ao plano Π? Tal ponto certamente *oclude* um ponto no solo. O ponto ocluso é dado pela intersecção do raio $\langle C, \mathbf{X}' \rangle$ com o plano Π. A Figura 3.3.1 ilustra a situação. A região que a Figura 3.3.1 exibe em azul marca as regiões em Π que são oclusas pelo corpo de um passageiro, no ângulo de visão da câmera exibida na Figura 3.3.1 (a). O mesmo efeito para outras duas câmeras é ilustrado na mesma figura.

A região em branco na Figura 3.3.1 (g) é de especial interesse. Ela se refere a uma região do plano Π que encontra-se encoberta em *todas* as câmeras. Havendo um único objeto em cena, a região em questão é completamente oclusa pelo mesmo objeto. Assumindo que este objeto seja, por exemplo, um cilindro, só há uma explicação possível: a região em questão corresponde a base do cilindro, sua *região de contato* com o solo. Logo, a região corresponde a localização do objeto no plano Π[2].

Quando há múltiplos objetos em cena, o que pode ser garantido é que a região no plano não pode ser vista diretamente por nenhuma das câmeras. Porém, quanto mais câmeras estiverem disponíveis, menores as chances disso ocorrer com pontos que não correspondam a regiões de contato dos vários objetos. No Capítulo 4 serão apresentados algoritmos que

[2]A afirmação obviamente não é válida para todo objeto. Um guada-sol poderia ocultar uma grande região de várias câmeras, embora apenas a ponta de seu cabo esteja em contato com o solo. No entanto, em relação ao problema de localização, a região totalmente oclusa pode ser aceita como o sítio do objeto em Π.

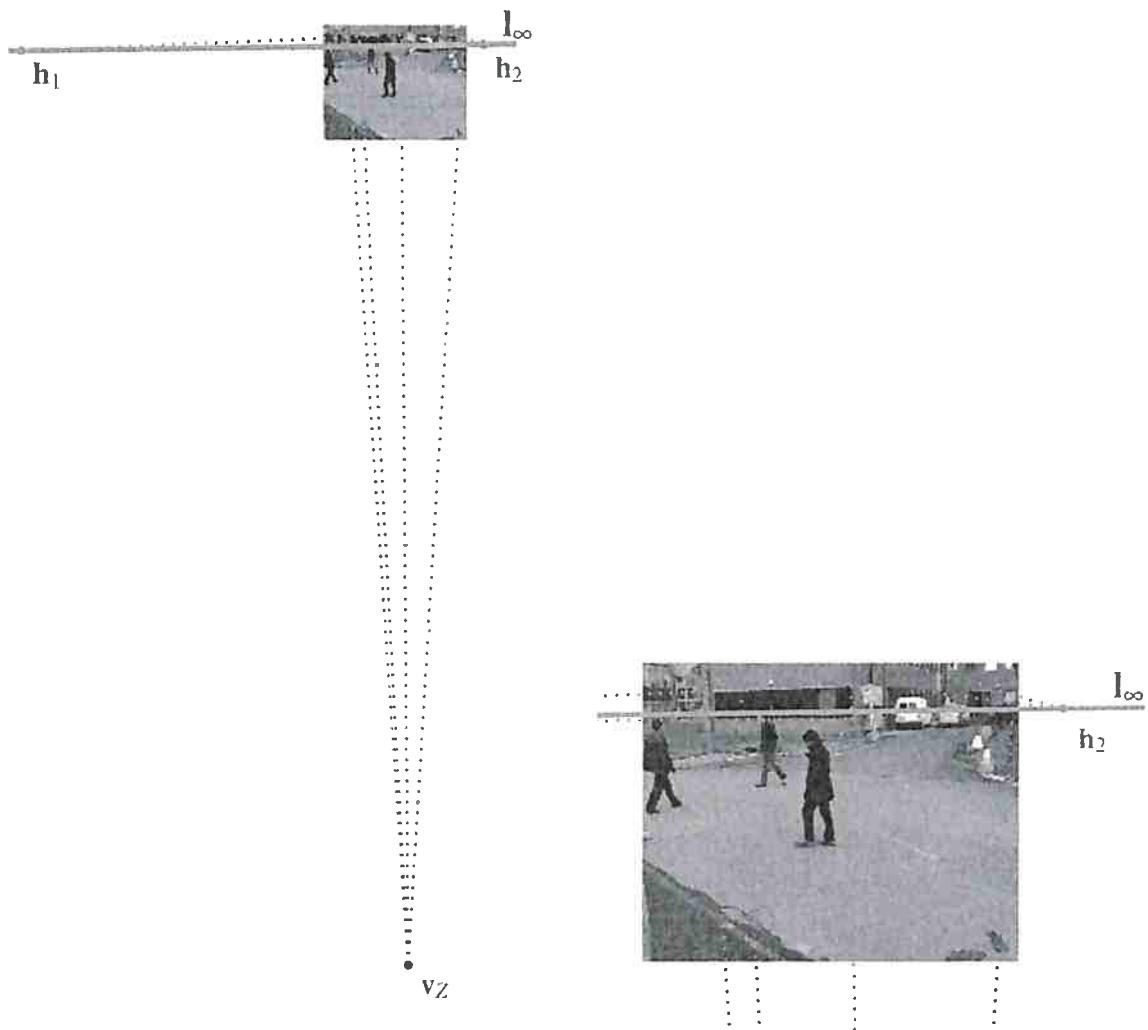


Figura 3.2.3: **Encontrando pontos afins**. O ponto afim $\mathbf{v}_Z$ pode ser obtido identificando duas linhas na imagem que correspondam a retas normais ao plano do solo. A linha do horizonte l_∞ é obtida de forma semelhante, identificando-se dois pares de linhas que correspondam a retas horizontais (paralelas ao plano do solo). As intersecções dessas retas produzem um par de pontos $\mathbf{h}_1$ e $\mathbf{h}_2$ que define a linha l_∞. (a) Imagem vista em uma escala que permite observar os pontos afins $\mathbf{v}_Z$, $\mathbf{h}_1$ e $\mathbf{h}_2$. (b) Detalhe: feixes de linhas correspondendo a retas verticais definem $\mathbf{v}_Z$, enquanto que feixes de linhas correspondendo a retas horizontais definem pontos de l_∞.

utilizam esta propriedade na localização de pessoas em cenas com múltiplas pessoas. Também será mostrado como algumas restrições, baseadas nas alturas dos indivíduos, podem ser aplicadas na tentativa de discernir regiões de contato entre as regiões totalmente oclusas.

A utilização desta propriedade na solução de problemas de localização e rastreamento com múltiplas câmeras é vista frequentemente na literatura recente. Mittal e Davis [38] apresentaram um conceito similar com visão estéreo no ano de 2003. Formulações mais elegantes que empregam homografias são vistas em 2006 nos trabalhos de Hu *et al.* [27], Kim e Davis [33] e Khan e Shah [31], que introduzem o termo *restrição homográfica* (*homographic constraint*). Novas soluções empregando restrição homográfica são apresentadas em 2008 por Santos e Morimoto [41], Eshel e Moses [17] e em 2009 novamente por Khan e Shah [32].

3.4 Estimação de homografias

A utilização da restrição homográfica requer que a matriz de homografia seja conhecida para cada câmera. Cada matriz de homografia $\mathtt{H}$ é uma matriz da forma

$$\mathtt{H} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix},$$

mas que só apresenta 8 graus de liberdade, por ser uma matriz homogênea. O fator de escala não importa, pois uma matriz $k\mathtt{H}$ representa a mesma transformação projetiva que uma matriz $\mathtt{H}$, se $k \neq 0$. Assim, só a *razão* entre os 9 componentes da matriz precisa ser determinada, definindo 8 graus de liberdade. Cada par de pontos correspondentes $\mathbf{X}_i \leftrightarrow \mathbf{x}_i$ restringe, através da transformação $\mathbf{X}_i = \mathtt{H}\mathbf{x}_i$ dois graus de liberdade, de forma que quatro correspondências definem completamente a matriz $\mathtt{H}$, que pode ser obtida através de um sistema linear.

Contudo, há dois problemas com este método mínimo. Para que o sistema linear possa ser resolvido, é necessário atribuir um valor $c \neq 0$ para uma das variáveis h_i e resolver o sistema linear para as oito variáveis restantes, obtendo uma solução $\tilde{\mathtt{H}}$ (uma opção comum

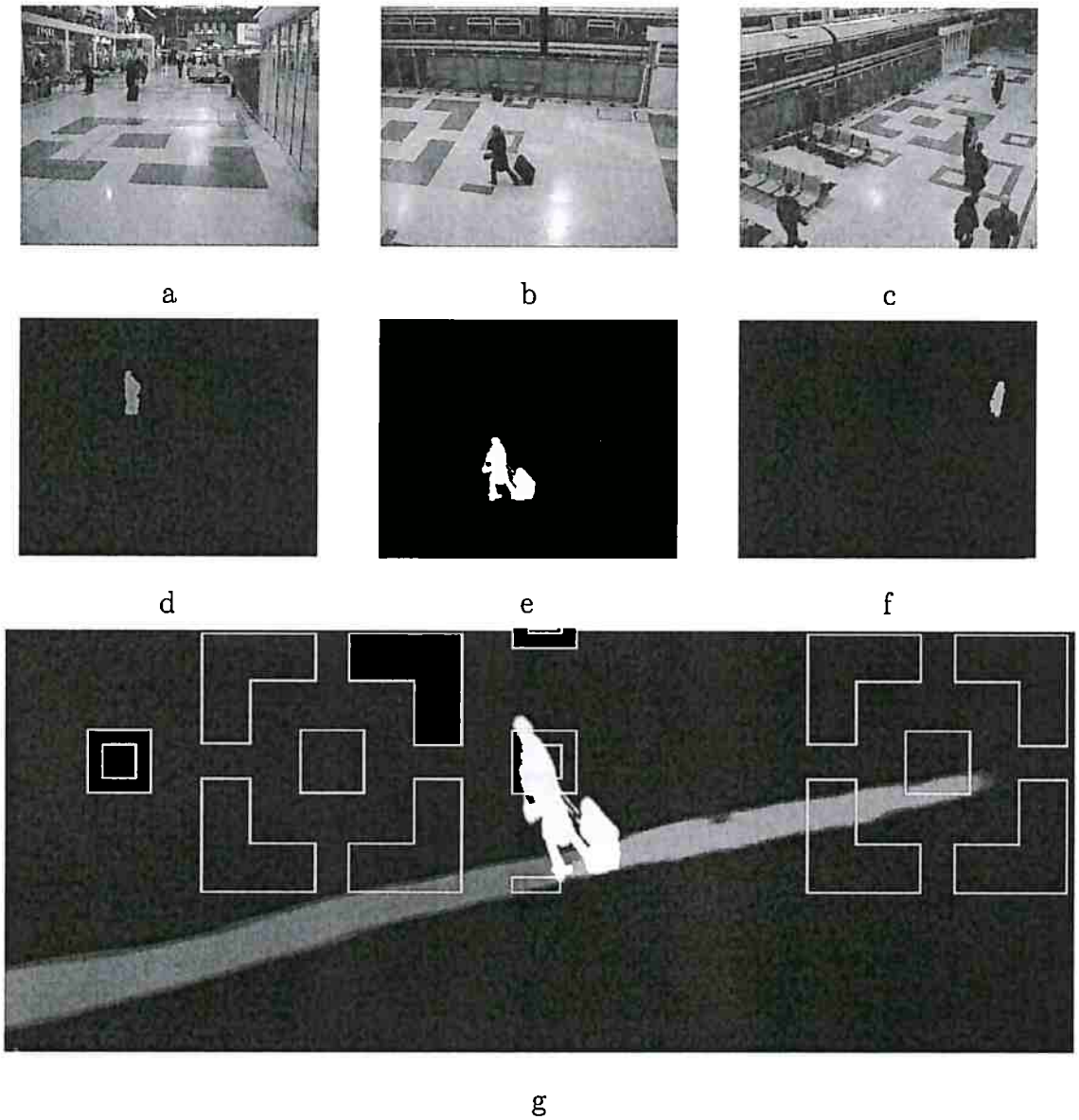


Figura 3.3.1: **Restrição homográfica**. Quadros obtidos por 3 câmeras na base de dados do PETS 2006 [48] são exibidos em (a), (b) e (c). Pixeis correspondentes à figura de uma mulher carregando uma mala de viagem são exibidos nas imagens (d), (e) e (f). Cada plano de imagem sofre uma transformação projetiva para o plano do solo Π. A imagem (g) exibe o plano Π com os resultados das homografias sobrepostos. A região branca demarca o ponto de contato dos objetos (pessoa e bagagem) com o solo. As regiões azul, verde e vermelha marcam, *para cada câmera*, quais as regiões no plano que estão oclusas pelos objetos.

é assumir $h_9 = 1$). O primeiro problema ocorre quando uma solução verdadeira para $\mathtt{H}$ apresenta $h_i = 0$, não existindo assim nenhum fator de escala k tal que $\mathtt{H} = k\tilde{\mathtt{H}}$. Tal situação ocorre, por exemplo, quando a origem do sistema de coordenadas em Π é mapeada para um ponto ideal no plano de imagem [24]. O método em questão produz resultados instáveis quando $h_i \approx 0$ e não é recomendado para uso geral.

O segundo problema é que, em geral, as correspondências fornecidas são inexatas. É desejável assim que mais do que quatro correspondências estejam disponíveis. O sistema fica então super-determinado e métodos de minimização são aplicados para a obtenção de um valor ótimo para $\mathtt{H}$.

Um método frequentemente empregado é a estimação homogênea [24]. A matriz $\mathtt{H}$ poder ser representada como um vetor da forma

$$\mathbf{h} = (h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8, h_9)^\top.$$

Considere n correspondências $\mathbf{X}_i \leftrightarrow \mathbf{x}_i$ disponíveis para a estimação, com $\mathbf{X}_i = (X_i, Y_i, 1)^\top$ e $\mathbf{x}_i = (x_i, y_i, 1)^\top$. A condição $\mathbf{X}_i = \mathtt{H}\mathbf{x}_i$ pode ser representada para as n correspondências através do produto

$$\mathtt{A}\mathbf{h} = \mathbf{0} \tag{3.14}$$

onde $\mathtt{A}$ é uma matriz $2n \times 9$ dada por

$$\mathtt{A} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1X_1 & -y_1X_1 & -X_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1Y_1 & -y_1Y_1 & -Y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2X_2 & -y_2X_2 & -X_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x_2Y_2 & -y_2Y_2 & -Y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & 1 & 0 & 0 & 0 & -x_nX_n & -y_nX_n & -X_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -x_nY_n & -y_nY_n & -Y_n \end{bmatrix}.$$

Para evitar a solução trivial, $\mathbf{h} = \mathbf{0}$, que não é de interesse, uma restrição deve ser associada a $\mathbf{h}$. A restrição $\|\mathbf{h}\| = 1$ pode ser imposta já que a escala de $\mathtt{H}$ não importa. O objetivo agora é minimizar a norma $\|\mathtt{A}\mathbf{h}\|$ sujeita a restrição $\|\mathbf{h}\| = 1$.

A matriz $\mathtt{A}$ pode ser decomposta em valores singulares na forma $\mathtt{A} = \mathtt{UDV}^\top$ (*singular value decomposition* – SVD [22]) . O objetivo agora é minimizar $\|\mathtt{UDV}^\top\mathbf{h}\|$. Como $\|\mathtt{UDV}^\top\mathbf{h}\| = \|\mathtt{DV}^\top\mathbf{h}\|$ e $\|\mathbf{h}\| = \|\mathtt{V}^\top\mathbf{h}\|$, pode-se alternativamente minimizar $\|\mathtt{DV}^\top\mathbf{h}\|$ sujeito a $\|\mathtt{V}^\top\mathbf{h}\| = 1$ como condição[3]. Aplicando-se uma substituição de variáveis $\mathbf{g} = \mathtt{V}^\top\mathbf{h}$, o problema pode ser redefinido como minimizar $\|\mathtt{Dg}\|$ sujeito à restrição $\|\mathbf{g}\| = 1$.

A matriz $\mathtt{D}$ é uma matriz diagonal 9×9 cujos elementos na diagonal principal são valores $\{d_i\}_{i=1..9}$ tais que $d_1 \geq d_2 \geq ... \geq d_9 > 0$. Segue daí que a solução do problema de minimização é dada por $\mathbf{g} = (0,0,0,0,0,0,0,0,1)^\top$. Como $\mathbf{h} = \mathtt{Vg}$, tem-se que $\mathbf{h}$ corresponde a última coluna de $\mathtt{V}$. A última coluna de $\mathtt{V}$ é ao auto-vetor de $\mathtt{A}^\top\mathtt{A}$ correspondente ao menor auto-valor.

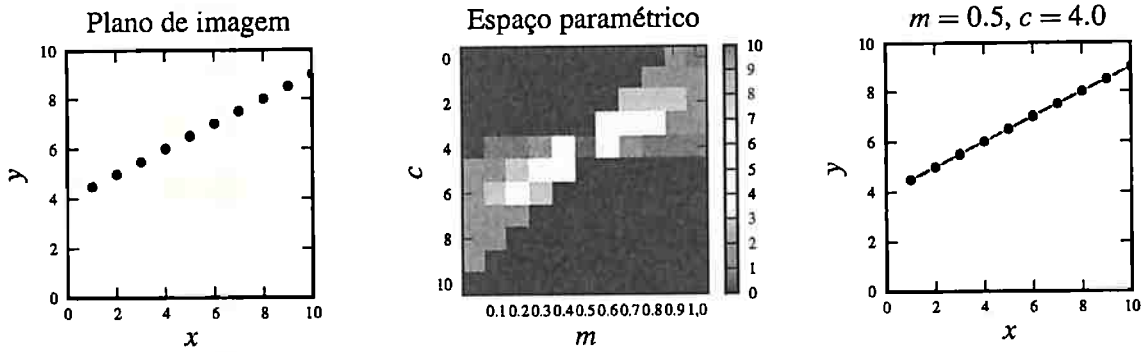
O método descrito acima é uma solução comumente utilizada na estimação de matrizes de homografia. Consiste em aplicar a decomposição SVD à matriz $\mathtt{A}$, obtendo-se assim a matriz $\mathtt{H}$ a partir da última coluna do vetor $\mathtt{V}$. Esta foi a solução adotada na obtenção de matrizes de homografia utilizadas neste trabalho. Hartley e Zisserman [24] abordam o problema de estimação de $\mathtt{H}$ com grande profundidade e o presente método é discutido por eles em detalhes. Os autores também dedicam um capítulo a várias outras alternativas de otimização existentes.

3.5 Transformada de Hough

A subtração de fundo pode ser utilizada para produzir imagens binárias representando as figuras dos objetos de interesse (Seção 3.1) enquanto que as homografias permitem relacionar essas figuras a regiões no plano do solo Π. Falta então um arcabouço que integre essas informações em um método de detecção de objetos.

A Transformada de Hough é um método bem conhecido para detecção de padrões complexos de pontos em imagens binárias [26, 28]. Um espaço compacto de parâmetros onde cada ponto representa um padrão de interesse é criado. Esse espaço é discretizado em um

[3] $\mathtt{U}^\top\mathtt{U} = \mathtt{V}^\top\mathtt{V} = \mathtt{I}$, onde $\mathtt{U}$ é uma matriz $2n \times 9$, $\mathtt{V}$ uma matriz 9×9 e $\mathtt{I}$ a matriz identidade 9×9.



a b c

Figura 3.5.1: **Transformada de Hough para detecção de linhas.** (a) Pontos marcados em um imagem binária – o objetivo é encontrar a reta que originou tais pontos. (b) Acumulador que representa retas da forma $y = mx + c$. Os valores de m foram discretizados no intervalo $[0, 1.0]$ enquanto que os valores de c encontram-se ente $[0, 10]$. É possível observar um valor máximo acumulado na célula $[m = 0.5,\ c = 4]$. Note que toda célula onde o acumulador não é vazio representa uma reta que intersecta ao menos um dos pontos observados. (c) Reta $y = 0.5x + 4$ exibida sobre o conjunto original de pontos.

conjunto de células acumuladoras . Cada observação no plano de imagem incrementa o valor de todas as células que representem um padrão que possa ser associado a tal observação.

O exemplo clássico utilizado é a detecção de linhas em imagens binárias (Figura 3.5.1). Assuma que um linha seja representada na forma $y = mx + c$. O espaço paramétrico é definido pelos valores de m e c, discretizados em uma grade (Figura 3.5.1 (b)). Para cada pixel (x, y), visitam-se todos os valores m_i. Para cada m_i, computa-se o valor c_i tal que $y = m_i x + c_i$, incrementando-se o valor da célula correspondente $[m_i, c_i]$. Máximos locais na grade de acumuladores observados correspondem às retas de interesse (Figura 3.5.1 (c)).

A transformada foi apresentada por Hough na forma de uma patente [26]. A aplicação estudada por Hough era a identificação de padrões em imagens obtidas por experimentos com câmaras de bolhas em Física. A transformada foi introduzida à comunidade de processamento de imagens por Rosenfeld [40]. Em um artigo bastante influente, Duda e Hart [14]

apresentaram uma variação do método de detecção de retas, utilizando coordenadas polares, e analisaram a eficiência do método e suas aplicações na identificação de outras curvas. O transformada generalizada de Hough, utilizada na detecção de formas arbitrárias, é apresentada por Ballard [1].

No próximo capítulo, algumas propriedades da Transformada de Hough serão discutidas e algoritmos para a detecção de objetos com o uso múltiplas câmeras serão apresentados.

Capítulo 4

Detecção de objetos por integração de suporte

Toda credibilidade, toda boa consciência, toda evidência de verdade vem apenas dos sentidos.

Friedrich Nietzsche - Além do bem e do mal

Os métodos apresentados neste capítulo têm como objetivo localizar vários indivíduos (ou objetos) em uma cena sobre um plano de referência. O plano em questão é de particular interesse: ele pode ser considerado uma *planta* do local. A vantagem de tal plano é que ele equivale a uma projeção ortográfica, como uma "tomada aérea" ou uma imagem de satélite, onde não há oclusão entre os objetos em trânsito e onde suas trajetórias podem ser facilmente estudadas, livre de efeitos de perspectiva. Com o uso de homografias, as localizações neste plano ideal podem ser mapeadas para o plano de imagem de qualquer câmera, permitindo sua localização em cada ângulo disponível, ou mesmo ângulos "virtuais" de câmeras sintéticas.

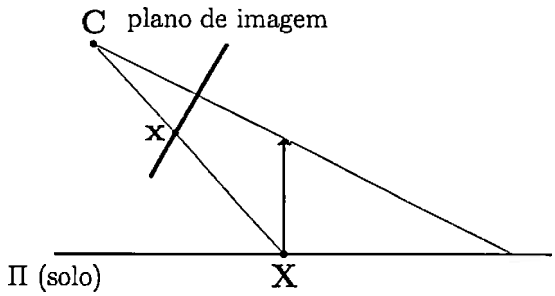
Este processo de detecção é baseado na Transformada de Hough. Esta transformada foi introduzida originalmente como um método para detecção de padrões complexos em imagens binárias. Padrões formados no plano de imagem são transformados para um espaço paramétrico, onde tais padrões podem ser representados de forma compacta. Converte-se assim um difícil problema de detecção no espaço da imagem em um problema mais tratável, a localização de máximos locais no espaço paramétrico [28].

Neste trabalho, o espaço paramétrico é o próprio plano de referência, que pode ser visto como uma *grade de acumuladores* dentro do arcabouço clássico da Transformada de Hough. Essa abordagem é similar ao conceito de *mapas de ocupação*, desenvolvido por Elfes [16] para navegação de robôs e recentemente usada por Fleuret *et al.* [19] para localização e rastreamento de indivíduos, como visto no Capítulo 2.

Um dos conceitos principais deste trabalho é o conceito de *suporte*, que é apresentado e desenvolvido na Seção 4.1. Essa idéia apresenta alguma similaridade com o método de projeção em histogramas, empregada pelo sistema W^4 de Haritaoglu *et al.* [23] e discutida no Capítulo 2. Porém, o suporte não é empregado na segmentação de indivíduos mas em sua localização, além de ser integrado, via restrição homográfica, em todas as câmeras, tratando situações de oclusão que não podem ser resolvida pelo método empregado pelo W^4. O arcabouço da Transformada de Hough permite uma formalização mais elegante do conceito de integração de suporte. O requisito essencial para o uso da transformada é a definição de um mapeamento entre os pontos da imagem e os parâmetros. Essa relação é definida na Seção 4.2. A eficiência da transformada e a qualidade dos resultados de detecção podem ser incrementados se for possível restringir o número de parâmetros endereçados [28]. A Seção 4.3 apresenta várias restrições que podem ser impostas à transformação caso sejam conhecidas algumas informações sobre a geometria da imagem, como pontos ideais e a linha do horizonte. A Seção 4.4 completa o arcabouço de Hough, definindo o método de detecção de máximos locais, utilizado para localizar regiões no plano com grande concentração de evidência. Finalmente, a Seção 4.6 apresenta um filtro capaz de remover possíveis falsos-positivos causados por casos difíceis de oclusão.

4.1 Suporte

Considere uma imagem binária F, resultante da subtração de fundo, onde pixeis são classificados como figura (proveniente de objetos de interesse) ou fundo, como discutido no Capítulo 3. O *suporte* de um ponto é a quantidade de figura vista acima dele. Trata-se de



a b

Figura 4.1.1: **Suporte em uma posição X.** (a) O suporte de um ponto $\mathbf{x}$ no plano de imagem é a quantidade de pixeis de figura vistos acima dele (destacados em amarelo na imagem). (b) Ele é uma especulação sobre a quantidade de "massa" de um objeto que pode estar se apoiando sobre a posição $\mathbf{X} = \mathtt{H}\mathbf{x}$ no plano do solo Π.

uma especulação sobre a quantidade de "massa" que pode estar se apoiando em um certo ponto do solo, baseada nos pixeis da imagem que compõem objetos.

Formalmente, seja Π o plano do solo e $\mathbf{x}$ um pixel correspondente à projeção de um ponto $\mathbf{X} \in \Pi$ no plano de imagem, isto é, $\mathbf{X} = \mathtt{H}\mathbf{x}$. Considere também o vetor $\vec{\mathbf{n}}_\mathbf{x}$, projeção do vetor normal de Π no plano de imagem em $\mathbf{x}$. O suporte $S(\mathbf{x})$ é o número de pixeis de figura vistos acima de $\mathbf{x}$, encontrados na semi-reta definida por $\mathbf{x}$ e $\vec{\mathbf{n}}_\mathbf{x}$. A Figura 4.1.1 ilustra o conceito.

$S(\mathbf{x})$ pode ser computado para qualquer $\mathbf{x} = \mathtt{H}^{-1}\mathbf{X}$, independentemente do fato de $\mathbf{X}$ corresponder ou não a verdadeira posição de um objeto em Π. Desta forma, o método especula sobre todas as possíveis localizações de objetos em cena, o que, combinado ao uso de múltiplos sensores, permite o tratamento de situações de oclusão. A ideia é ilustrada na Figura 4.1.2. O único indivíduo visto em na Figura 4.1.2 (a) produzirá algum suporte para diferentes localizações no plano. Contudo, pontos mais próximos a real localização do objeto apresentarão maior suporte. No exemplo da figura, $S(\mathbf{x}) > S(\mathbf{y}) > S(\mathbf{z})$. No caso de oclusão apresentado na Figura 4.1.2 (c), a região ao redor das localizações dos três indivíduos vistos

apresentarão maior suporte.

4.1.1 Orientação do plano

Para determinar quais pixeis estão "acima" de uma certa posição, utiliza-se o vetor normal $\vec{\mathbf{n}}_{\mathbf{x}}$, a projeção do vetor normal de Π no plano de imagem na posição $\mathbf{x}$. Como visto no Capítulo 3, a direção vertical é representada pelo ponto ideal $\mathbf{v}_Z$. Conhecendo-se $\mathbf{v}_Z$, a projeção do vetor normal em $\mathbf{x}$ pode ser obtida por

$$\vec{\mathbf{n}}_{\mathbf{x}} = u \cdot \frac{\mathbf{v}_Z - \mathbf{x}}{\|\mathbf{v}_Z - \mathbf{x}\|}, \tag{4.1}$$

onde u é uma constante que assume o valor 1 ou -1 dependendo da orientação da imagem, que precisa ser informada ao sistema[1]. A Figura 4.1.3 ilustra como o vetor $\vec{\mathbf{n}}_{\mathbf{x}}$ varia com a posição $\mathbf{x}$ no plano de imagem. No caso do plano de imagem ser ortogonal ao plano Π, o ponto $\mathbf{v}_Z$ será um ponto ideal (um ponto no infinito). Nesse caso, todas as projeções $\vec{\mathbf{n}}_{\mathbf{x}}$ serão iguais.

4.1.2 Compensação de efeitos de perspectiva

Há um problema com o simples acúmulo de pixeis. Transformações projetivas associadas à aquisição de imagens tornam menores os objetos mais afastados da câmera no plano de projeção. Este fato bem conhecido é ilustrado na Figura 4.1.4. A figura apresenta, através de sobreposição de imagens, a mesma pessoa vista em duas posições diferentes. Conforme se afasta da câmera, o tamanho *em pixeis* do indivíduo diminui. Isto significa que um objeto de altura h apresentará, no plano de imagem, diferentes alturas em pixeis, dependendo da posição do objeto em relação à câmera.

Considere um objeto de altura h_r visto pela câmera na posição $\mathbf{r}_\perp$, que corresponde a uma posição $\mathbf{R} = \mathtt{H}\mathbf{r}_\perp$ no plano Π. É possível computar o valor $\mathfrak{h}_r(\mathbf{x})$ correspondente à altura

[1] A orientação real de uma imagem só é sabida de fato por pessoas com conhecimento sobre o posicionamento da câmera. Um fotógrafo, por exemplo, pode iludir sua audiência invertendo o posicionamento da câmera e removendo referenciais de orientação.

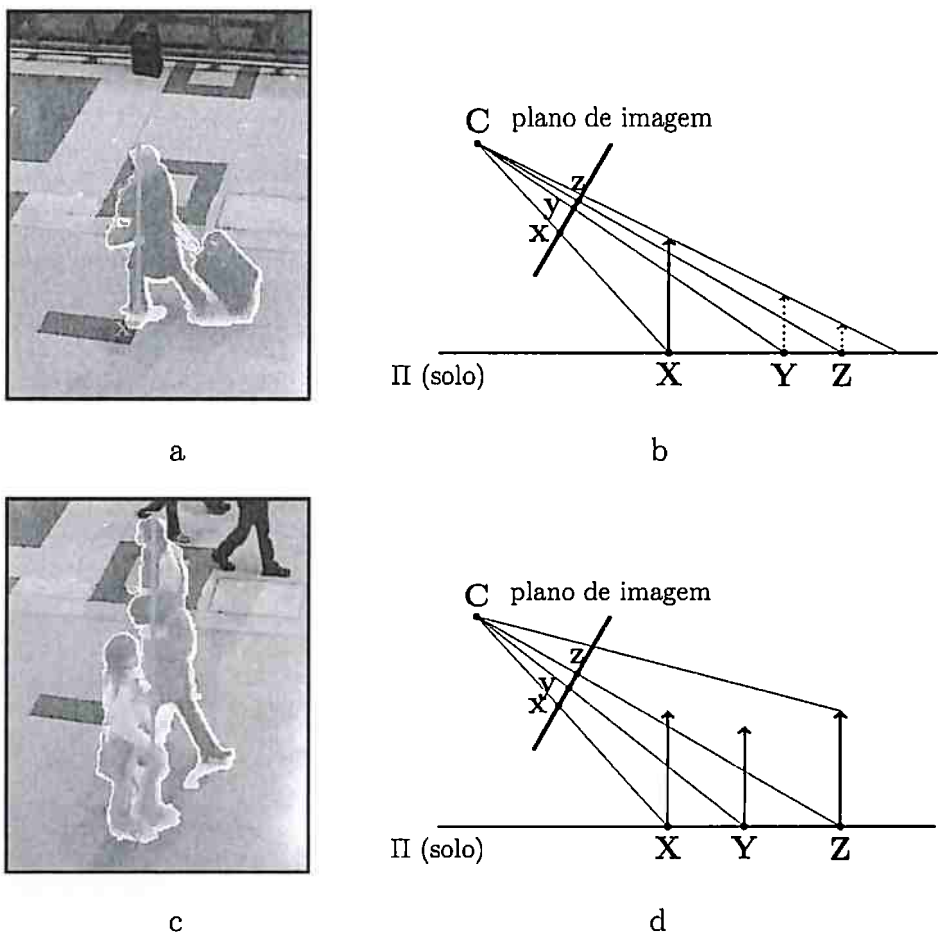


Figura 4.1.2: **Avaliação de sítios utilizando suporte**. O suporte pode ser computado para as várias localizações possíveis em no plano Π. Valores elevados de suporte deveriam ser encontrados próximos aos sítios onde os objetos se localizam, como visto nas posições **X**, **Y** e **Z** em (d) e na posição **X** em (b). Baixos valores indicam ausência de objetos, como é o caso das posições **Y** e **Z** em (b). Regiões amarelas em (a) e (c) indicam figuras. Linhas cheias em (b) e (d) indicam posições reais de objeto.

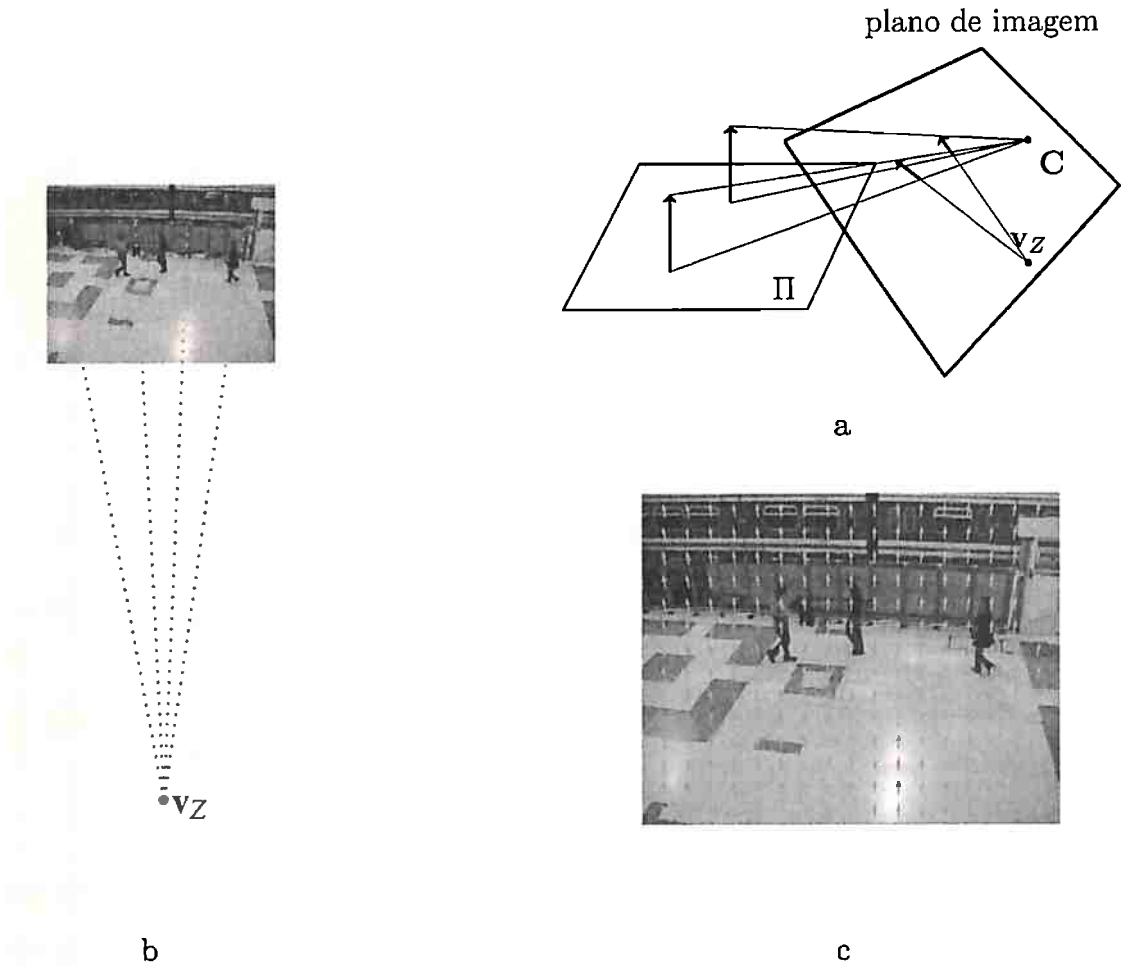


Figura 4.1.3: **Orientação e vetor normal.** A direção do vetor normal de Π (a) é representada pelo ponto ideal $\mathbf{v}_Z$ (b). Uma vez informada a orientação da normal no plano de imagem, os vetores $\vec{\mathbf{n}}_\mathbf{x}$ podem ser determinados para cada ponto $\mathbf{x}$ (c).

em pixeis que o objeto apresentaria se fosse encontrado na posição $\mathbf{X} = \mathtt{H}\mathbf{x}$ em Π, para todo pixel $\mathbf{x}$ no plano de imagem. Em outras palavras, é possível antecipar a altura em pixeis correspondente a h_r que o objeto apresentaria na imagem se deslocado à posição $\mathbf{X}$ no solo.

Seja $\mathbf{r}_\top$ a projeção de $\mathbf{r}_\perp$ em um plano paralelo a uma distância h_r de Π, como visto na Figura 4.1.4. Considere $d(\mathbf{x}, \mathbf{y})$ como a distância *em pixeis* entre dois pontos $\mathbf{x}$ e $\mathbf{y}$ no plano de imagem e assuma que $d(\mathbf{r}_\perp, \mathbf{r}_\top)$ é conhecida (a altura de referência). Assim, a altura $d(\mathbf{x}_\perp, \mathbf{x}_\top)$ do objeto quando este se encontra na posição $\mathbf{x}_\perp$ pode ser obtida utilizando-se a invariância da *razão cruzada* sob transformações projetivas.

Criminisi *et al.* [11] aplicaram a invariância da razão-cruzada para obter a relação

$$\frac{h_r}{h_q} = 1 - \frac{d(\mathbf{r}_\top, \mathbf{c}_\mathbf{r})d(\mathbf{r}_\perp, \mathbf{v}_Z)}{d(\mathbf{r}_\perp, \mathbf{c}_\mathbf{r})d(\mathbf{r}_\top, \mathbf{v}_Z)} \tag{4.2}$$

entre a altura de referência h_r e a altura da câmera h_q (a distância do centro de projeção da câmera q ao plano Π) quando o objeto de referência está localizado em $\mathbf{r}_\perp$. O ponto $\mathbf{c}_\mathbf{r}$ é a imagem de um ponto a uma distância h_q do plano Π, logo pertencente a $\mathbf{l}_\infty$, a projeção da linha no infinito de Π no plano de imagem de q (Figura 4.1.4).

Uma equação similar pode ser obtida quando o objeto de referência se encontra em $\mathbf{x}_\perp$

$$\frac{h_r}{h_q} = 1 - \frac{d(\mathbf{x}_\top, \mathbf{c}_\mathbf{x})d(\mathbf{x}_\perp, \mathbf{v}_Z)}{d(\mathbf{x}_\perp, \mathbf{c}_\mathbf{x})d(\mathbf{x}_\top, \mathbf{v}_Z)}. \tag{4.3}$$

Considere $a(\mathbf{x}_\perp) = d(\mathbf{x}_\perp, \mathbf{v}_Z)$ e $b(\mathbf{x}_\perp) = d(\mathbf{x}_\perp, \mathbf{c}_\mathbf{x})$. Assim, os termos em $\mathbf{x}_\top$ podem ser escritos como

$$d(\mathbf{x}_\top, \mathbf{v}_Z) = a(\mathbf{x}_i^q) - \mathfrak{h}_r(\mathbf{x}_\perp) \tag{4.4}$$

$$d(\mathbf{x}_\top, \mathbf{c}_\mathbf{x}) = b(\mathbf{x}_i^q) - \mathfrak{h}_r(\mathbf{x}_\perp). \tag{4.5}$$

Seja

$$c = \frac{d(\mathbf{r}_\top, \mathbf{c}_\mathbf{r})d(\mathbf{r}_\perp, \mathbf{v}_Z)}{d(\mathbf{r}_\perp, \mathbf{c}_\mathbf{r})d(\mathbf{r}_\top, \mathbf{v}_Z)}. \tag{4.6}$$

Aplicando a igualdade entre as Equações 4.2 e 4.3, após algumas manipulações algébricas simples obtém-se:

$$\mathfrak{h}_r(\mathbf{x}_\perp) = \frac{a(\mathbf{x}_\perp)b(\mathbf{x}_\perp)(1-c)}{a(\mathbf{x}_\perp) - b(\mathbf{x}_\perp)c} \tag{4.7}$$

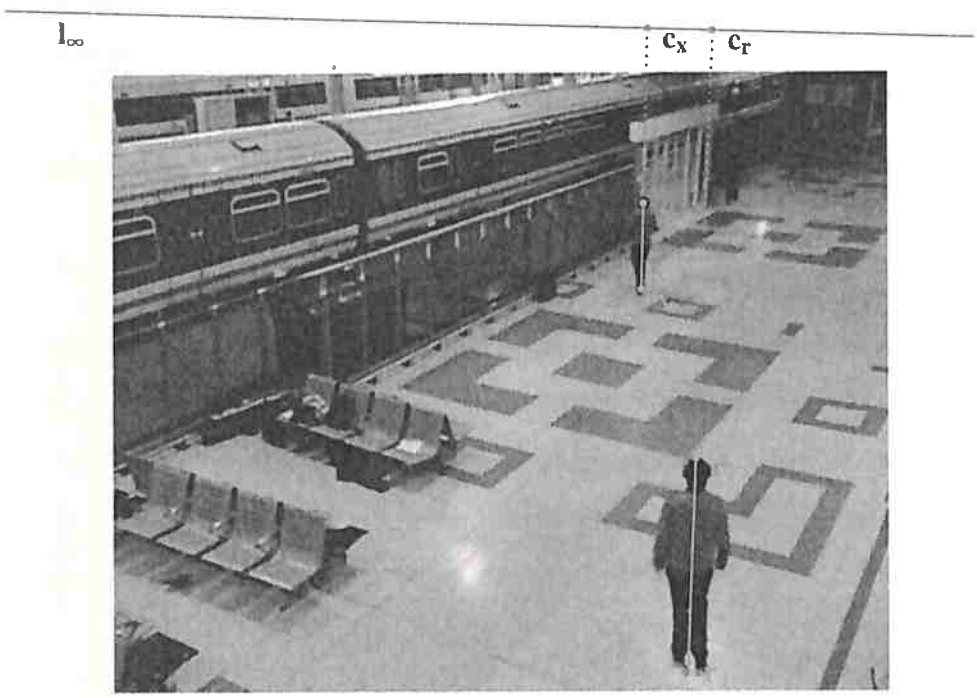


Figura 4.1.4: **Uso da razão cruzada para compensação de efeitos de perspectiva.** A figura exibe uma sobreposição de imagens que mostra o mesmo indivíduo em duas posições diferentes. A razão cruzada pode ser usada para estabelecer o tamanho, em pixeis, que um objeto de altura h_r apresentaria em posições diferentes do plano (vide texto para maiores detalhes).

Logo, o valor de $\mathfrak{h}_r(\mathbf{x})$ pode ser visto como uma função de $\mathbf{x}$ e pré-computado para referência. Se usado como fator de normalização, pode compensar efeitos de perspectiva, como será visto a seguir.

4.1.3 Restrições de altura para os objetos

Devido à oclusão, aglomerados de pixeis de figura poderiam ser formados por diversos objetos. Aglomerados grandes e alongados, formados quando os indivíduos encontram-se alinhados verticalmente, cria uma região de alto suporte, como pode ser observado na Figura 4.1.5 (a e b). Como ilustrado na figura, a quantidade de suporte observada pode ser erroneamente interpretada como a existência de um grande objeto na base da região.

Considere o pixel $\mathbf{x}$ no plano de imagem, pertencente a um indivíduo qualquer. Seja $\mathbf{x}_\perp$ a localização do objeto responsável por $\mathbf{x}$, isto é, o ponto de contato na reta $\langle \mathbf{x}, \mathbf{v}_Z^q \rangle$ (eventualmente, $\mathbf{x} = \mathbf{x}_\perp$). Se $h_{\max}$ é a altura do maior objeto esperado na cena, então a maior distância possível, em pixeis, entre $\mathbf{x}$ e $\mathbf{x}_\perp$ é $\mathfrak{h}_{\max}(\mathbf{x}_\perp)$. O valor de $\mathfrak{h}_{\max}(\cdot)$ é computado de acordo com as Equações 4.7 e 4.6, sendo c obtido através de pontos de referência $\mathbf{r}_\perp$ e $\mathbf{r}_\top$ relativos a um objeto de altura $h_{\max}$. Essa é a situação na qual $\mathbf{x}$ é o ponto extremo de um objeto, $\mathbf{x}_\top$ (a "cabeça" de um indivíduo, por exemplo) e o objeto apresenta a altura máxima permitida pelo modelo.

Considerando essa restrição à altura e a orientação da imagem, é possível obter um algoritmo eficiente para o cômputo do suporte.

4.1.4 Algoritmo básico

O algoritmo descrito a seguir foi introduzido no *XXI Brazilian Symposium on Computer Graphics and Image Processing* (SIBGRAPI 2008) [41] (uma versão estendida do artigo [42] foi submetida à *Pattern Recognition Letters*, onde se encontra, até o presente momento, em processo revisão).

Seja $\langle \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \rangle$ um segmento de pixeis, obtido restringindo-se a linha $\langle \mathbf{x}_1, \mathbf{v}_Z \rangle$ pelos limites da imagem, como ilustrado na Figura 4.1.6. O Algoritmo 1 computa o suporte pela

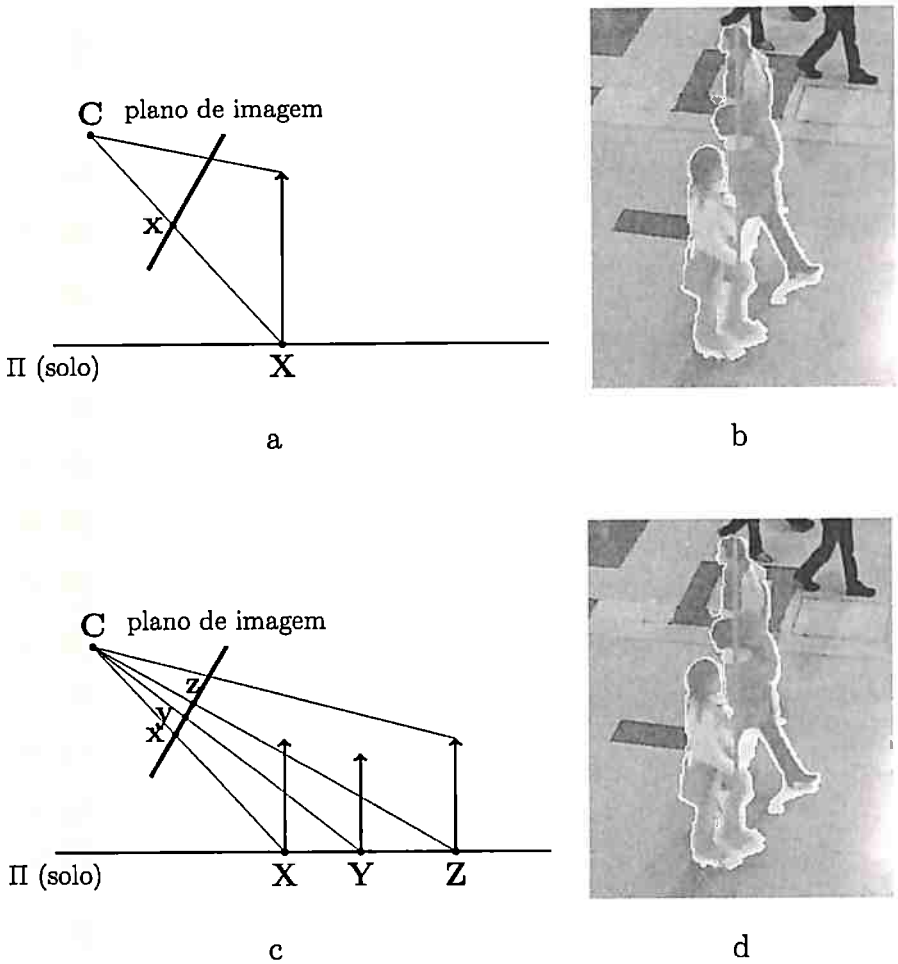


Figura 4.1.5: **Restrições para a altura máxima de um objeto**. Oclusão entre objetos pode produzir grandes aglomerados alongados de figura (b). Tal evento poderia ser erroneamente interpretado como a presença de um grande objeto na cena (a). Ao restringir-se a altura máxima de um objeto, cria-se uma região de candidatos, todos com quantidades similares de suporte, como o intervalo entre **X** e **Z** em (c). Integrando suporte proveniente de outras câmeras, as posições **X**, **Y** e **Z** podem ser discriminadas.

contagem do número de pixeis de figura projetados sobre cada $\mathbf{x}_i$, utilizando $\mathfrak{h}_{max}(\mathbf{x}_i)$ como fator de normalização.

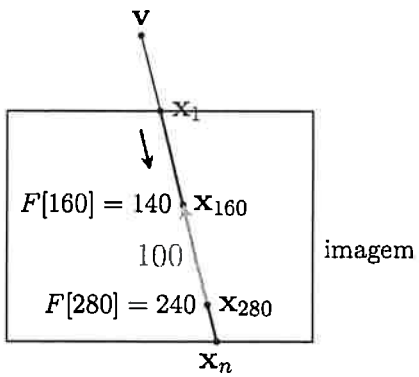


Figura 4.1.6: **Iteração do algoritmo básico para suporte**. Uma iteração do Algoritmo 1 para $i = 280$. O código na Linha 8 inspeciona o pixel $\mathbf{x}_{160}$, o qual corresponderia à cabeça do mais altor objeto esperado. Dado que $F[160] = 140$, há então 100 pixeis de figura entre $\mathbf{x}_{160}$ e $\mathbf{x}_{280}$.

Para ilustrar a operação do algoritmo, considere o pixel $\mathbf{x}_{280}$ e a situação representa na Figura 4.1.6. Há 240 pixeis existentes acima dessa localização, seguindo a direção normal – esta informação encontra-se armazenada no contador $F[280]$. Como $\mathfrak{h}_{max}(\mathbf{x}_{280}) = 120$, sabe-se que o mais alto objeto esperado, se encontrado na posição $\mathbf{x}_{280}$, cobriria 120 pixeis na direção normal, atingindo o pixel $\mathbf{x}_{160}$ (linha 8 do algoritmo). Dado que $F[160] = 140$ (há 140 pixeis marcados como figura observados acima de $\mathbf{x}_{160}$), há 100 pixeis entre $\mathbf{x}_{280}$ e $\mathbf{x}_{160}$ que pertencem a objetos. Este número, normalizado por $\mathfrak{h}_{max}(\mathbf{x}_{280})$ é o suporte evidenciado em $\mathbf{x}_{280}$.

O algoritmo de Bresenham [7] para determinação de uma linha em uma grade de pixeis pode ser utilizado, em conjunto com o ponto ideal $\mathbf{v}_Z$, para determinar todos os segmentos de entrada e computar o suporte de modo eficiente para todos os pixeis do plano de imagem de uma câmera.

Finalmente, a restrição homográfica pode ser empregada para integrar o suporte obtido por várias câmeras diferentes. Seja $S^q(\cdot)$ o suporte obtido para a q-ésima câmera. Todo o

Algoritmo 1 Algoritmo para cômputo do suporte $S(\mathbf{x}_i)$ para todos os pixeis $\mathbf{x}_i$ pertencentes a um segmento de entrada. O algoritmo assume que os pixeis no segmento estão ordenados em sentido inverso ao da normal do plano Π. F corresponde ao conjunto pixeis classificados como figura.

1: **procedimento** SUPORTE($\langle \mathbf{x}_1, ..., \mathbf{x}_n \rangle$, F, $\mathfrak{h}_{\max}$, $\mathfrak{h}_{\min}$)
2: $C[0] \leftarrow 0$
3: **para** $i \leftarrow 1, n$ **faça**
4: **se** $\mathbf{x}_i \in F$ **então**
5: $C[i] \leftarrow C[i-1] + 1$
6: **senão**
7: $C[i] \leftarrow C[i-1]$
8: $j \leftarrow i - \mathfrak{h}_{\max}(\mathbf{x}_i)$
9: **se** $j > 0$ **então**
10: $h \leftarrow (C[i] - C[j])/\mathfrak{h}_{\max}(\mathbf{x}_i)$
11: **senão**
12: $h \leftarrow C[i]/\mathfrak{h}_{\max}(\mathbf{x}_i^q)$
13: **se** $h \geq \mathfrak{h}_{\min}(\mathbf{x}_i)/\mathfrak{h}_{\max}(\mathbf{x}_i)$ **então**
14: $S(\mathbf{x}_i) \leftarrow h$
15: **senão**
16: $S(\mathbf{x}_i) \leftarrow 0$
17: **return** S_q

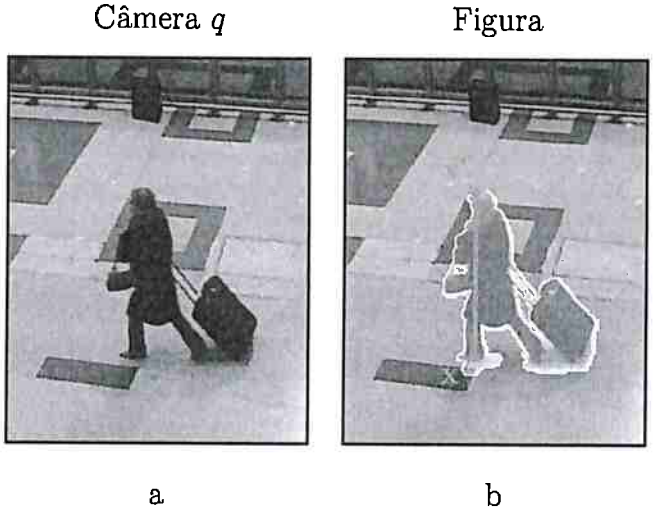


Figura 4.1.7: **Suporte utilizado na estimação da altura.** A distância $d(\mathbf{x}_\perp, \mathbf{x}_\top)$ pode ser utilizada para calcular a altura de um objeto [11]. Como o suporte é uma aproximação do valor de $d(\mathbf{x}_\perp, \mathbf{x}_\top)$, ele pode ser utilizado para estimação da altura dos objetos identificados.

suporte observado pode ser integrado em Π por

$$A[\mathbf{X}] = \sum_{q=1}^{Q} S^q(\mathrm{H}^{q-1}\mathbf{X}). \tag{4.8}$$

4.1.5 Suporte como estimativa da altura de um objeto

Há outra vantagem no uso de $\mathfrak{h}_{\max}(\cdot)$ para a compensação de efeitos de perspectiva. Os valores encontrados no acumulador A, se normalizados pelo números de câmeras capazes de observar cada posição, são uma *aproximação da altura* do objeto. Este valor é definido em relação à altura $h_{\max}$.

No caso de pessoas, pode-se assumir que os indivíduos vistos em cena encontram-se em posição ereta - com cabeça, coluna e bacia alinhados. Está hipótese é razoável para indivíduos caminhando em locais públicos [2]. Considere a Figura 4.1.7. A altura do indivíduo pode ser estimada através da distância entre os pontos $\mathbf{x}_\perp$ ("pés") e $\mathbf{x}_\top$ ("cabeça"). Como $\mathfrak{h}_{\max}(\mathbf{x}_\perp)$ é

[2]Tal hipótese pode não ser válida para pessoas correndo, quando é comum o indivíduo projetar cabeça e tronco à frente.

a altura em pixeis esperada para o objeto de altura $h_{\max}$ se posicionado em $\mathbf{x}_\perp$ e como $S(\mathbf{x}_\perp)$ é determinado com base no número de pixeis de figura observados acima de $\mathbf{x}_\perp$, tem-se que

$$S(\mathbf{x}\perp) \approx \frac{d(\mathbf{x}_\perp, \mathbf{x}_\top)}{\mathfrak{h}_{\max}(\mathbf{x}_\perp)} \tag{4.9}$$

Seja h a altura do indivíduo observado em posição $\mathbf{x}_\perp$. Embora transformações projetivas não preservem as razão entre comprimentos de segmentos, a razão cruzada pode ser aplicada para estabelecer que [24](pp. 222)

$$\frac{d(\mathbf{x}_\perp, \mathbf{x}_\top)[d(\mathbf{x}_\perp, \mathbf{v}_Z) - \mathfrak{h}_{\max}(\mathbf{x}_\perp)]}{\mathfrak{h}_{\max}(\mathbf{x}_\perp)[d(\mathbf{x}_\perp, \mathbf{v}_Z) - d(\mathbf{x}_\perp, \mathbf{x}_\top)]} = \frac{h}{h_{\max}}. \tag{4.10}$$

Combinando as duas equações, tem-se

$$S(\mathbf{x}\perp)\frac{d(\mathbf{x}_\perp, \mathbf{v}_Z) - \mathfrak{h}_{\max}(\mathbf{x}_\perp)}{d(\mathbf{x}_\perp, \mathbf{v}_Z) - d(\mathbf{x}_\perp, \mathbf{x}_\top)} \approx \frac{h}{\mathfrak{h}_{\max}}. \tag{4.11}$$

Particularmente quando $d(\mathbf{x}_\perp, \mathbf{v}_Z) \gg \mathfrak{h}_{\max}(\mathbf{x}_\perp) > d(\mathbf{x}_\perp, \mathbf{x}_\top)$,

$$S(\mathbf{x}\perp) \approx \frac{h}{\mathfrak{h}_{\max}}. \tag{4.12}$$

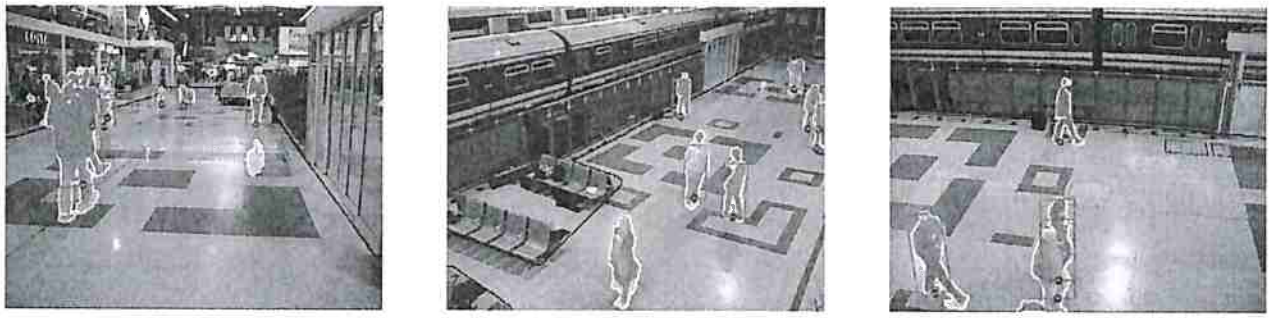
Seja $\mathfrak{v}(\mathbf{X})$ o número de câmeras que têm o ponto $\mathbf{X} \in \Pi$ dentro de seu campo de visão, dentre as Q câmeras disponíveis. Se $h/h_{\max}$ for estimado como uma média das aproximações obtidas por cada câmera, tem-se

$$\frac{A[\mathbf{X}_\perp]}{\mathfrak{v}(\mathbf{X}_\perp)} \approx \frac{h}{h_{\max}}. \tag{4.13}$$

Quando não há oclusão e a subtração de fundo não gera falsos-positivos, é mais comum observar $S(\mathbf{x}_\perp) \leq d(\mathbf{x}_\perp, \mathbf{x}_\top)$. Isto se deve principalmente a erros de subtração de fundo (camuflagem) e a situações onde certas partes do corpo do indivíduo não estejam alinhadas com o tronco, mais comumente as pernas. Para este caso, $\frac{A[\mathbf{X}_\perp]}{\mathfrak{v}(\mathbf{X}_\perp)} \lessapprox \frac{h}{h_{\max}}$. Esta é a situação mais comum quando a densidade de objetos em cena é baixa.

Quando há oclusão ou quando a subtração de fundo produzir falsos-positivos (evento comum com sombras, por exemplo), o suporte de uma localização pode receber pixeis provenientes de vários objetos diferentes ou de falsos-positivos. Neste caso $\frac{h}{h_{\max}} \lessapprox \frac{A[\mathbf{X}_\perp]}{\mathfrak{v}(\mathbf{X}_\perp)} \leq 1$.

A Figura 4.1.8 mostra o resultado obtido ao utilizar-se o suporte como estimador de altura. Os retângulos em vermelho ao redor de cada indivíduo são determinados pela posição



a b c

Figura 4.1.8: **Estimação da altura de indivíduos.** A figura mostra o resultado de detecção e estimação de altura utilizando-se o acumulador A. As posições correspondem a máximos locais encontrados em A, transformados de volta ao plano de imagem por homografia. A posição, a direção normal e a altura estimada são utilizadas para definir um retângulo ajustado ao indivíduo. Quando há falhas na subtração de fundo, vistas sobretudo em (a), a altura é sub-estimada os retângulos não se ajustam bem às silhuetas individuais - o que pode ser observado para os indivíduos no centro de (b). O mesmo ocorre quando as pernas não estão alinhadas com o tronco, como observado para o indivíduo na parte superior da figura em (c).

$\mathbf{x}_\perp$ (máximo local em A), pela direção normal e pela altura relativa encontrada. As regiões destacadas em amarelo representam o resultado da subtração de fundo. É possível observar retângulos bem ajustados à silhueta do indivíduo quando a subtração de fundo recupera as figuras corretamente. Havendo falsos-negativos ou quando a pessoa não está perfeitamente alinhada, a altura é sub-estimada e os retângulos não conseguem conter a cabeça do indivíduo. De maneira geral, a estimação é útil e pode ser empregada em algumas aplicações.

4.2 Suporte e a Transformada de Hough

O método descrito na seção anterior é eficaz na localização de indivíduos. O conceito de suporte assume cada pixel de figura como uma *evidência* da presença de um ou mais objetos em várias posições possíveis. A compreensão do processo pelo qual um pixel evidencia certos

sítios no plano de imagem e de como o conhecimento sobre a geometria da cena pode ser empregado para restringir o número de sítios pode ser auxiliada com o uso do arcabouço da Transformada de Hough, apresentada no Capítulo 3.

A Transformada de Hough pode ser vista como um processo de acúmulo de evidência [28]. Os pixeis classificados como figura através da subtração de fundo são utilizados como evidências da presença de um ou vários objetos. Esta evidência é transformada para o espaço paramétrico onde ela pode ser integrada e analisada. Nesse espaço, sítios que acumulam a maior quantidade de evidência são os melhores candidatos à localização de objetos. Se as evidências são provenientes de vários sensores diferentes, maior a confiança no sítio como sendo a localização de um objeto.

Ajustamento de protótipos (*template matching*) e a Transformada de Hough são fortemente relacionados. A diferença principal entre os métodos é que o ajustamento de protótipos varre todo o espaço da imagem, comparando os protótipos às observações encontradas em todos os sítios possíveis. O número de protótipos testados pode ser muito grande, devido a uma explosão combinatória de observações. A Transformada de Hough quantiza o espaço de parâmetros de forma que, se o número de observações for suficientemente maior que esse espaço, a transformada se torna uma alternativa mais eficiente [14].

Seja $\mathbf{x}$ um ponto de figura encontrado no plano de imagem da câmera q. A relação que define a Transformada de Hough em questão é dada por

$$f_q(\mathbf{X}_i, \mathbf{x}) = \mathbf{X}_i \cdot (\mathtt{H}^q\mathbf{x} \times \mathbf{V}_Z^q) = 0 \tag{4.14}$$

onde $\mathbf{V}_Z^q = \mathtt{H}^q\mathbf{v}_Z^q$. Em outras palavras, os parâmetros associados a $\mathbf{x}$ são todos os pontos $\mathbf{X}_i \in \Pi$ que se encontram na linha definida por $\mathtt{H}^q\mathbf{x}$ e $\mathbf{V}_Z^q$. A interpretação física é simples: *não havendo informação sobre a orientação da imagem q*, o ponto de figura $\mathbf{x}$ pode ter sido gerado por um objeto localizado em qualquer um dos pontos $\mathbf{X}_i$ dados pela Equação 4.14.

4.2.1 Integração de múltiplos sensores

A transformação oriunda da relação definida pela Equação 4.14 pode ser realizada para todas as Q câmeras disponíveis. É pelo acúmulo da evidência proveniente de várias câmeras que as reais localizações podem ser identificadas em situações de oclusão como as ilustradas na Figura 4.1.2 e na Figura 2.2.1, vista anteriormente no Capítulo 2.

Os pontos $\mathbf{X}_i$ do espaço paramétrico que apresentam valores diferentes de zero em seus acumuladores podem ser divididos em duas categorias. Alguns deles são *pontos de contato*, posições em Π onde objetos tocam o solo, sendo assim as localizações desejadas. Os pontos restantes são *pontos encobertos*, isto é, regiões em Π que não podem ser vistas diretamente pela câmera, pois estão oclusas pelos objetos na cena. Esta é a condição dos pontos $\mathbf{Y}$ e $\mathbf{Z}$ vistos na Figura 4.1.2 (d).

O que faz o método efetivo na localização de objetos é o fato dos *pontos de contato serem consistentes entre câmeras.* Tal fato se deve à restrição homográfica apresentada nos Capítulos 2 e 3. Como visto anteriormente, pontos encobertos não gozam dessa propriedade. Descartando-se falhas na subtração de fundo, a única forma de um ponto encoberto acumular tanta evidência quanto um ponto de contato é se ele estiver encoberto por algum objeto em *todas* as câmeras disponíveis, como ilustrado na Figura 4.2.1. Esta é um característica interessante do sistema pois significa que a localização em questão não está visível por *nenhum* dos sensores. A decisão sobre a existência de um objeto neste local pode ser postergada, aguardando informação adicional proveniente, por exemplo, de um módulo de rastreamento (o objeto-alvo se encontra oculto no momento). Outra alternativa é discriminar a situação e descartar o possível falso-positivo, como será mostrado Seção 4.6.

4.2.2 Um algoritmo simples pela Transformada de Hough

O Algoritmo 2 é um algoritmo simples para o cômputo do suporte pela Transformada de Hough, através da relação definida na Equação 4.14.

A Figura 4.2.2 exibe os resultados da execução do algoritmo para um instante da sequência S07 da base de dados PETS 2006. É possível identificar máximos locais correspondentes

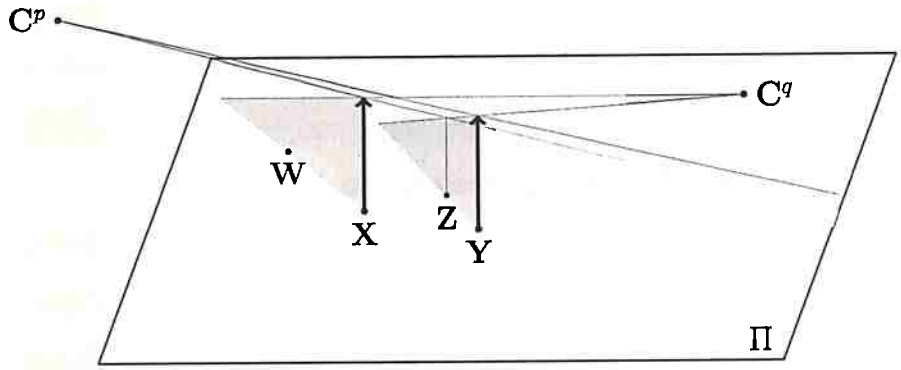


Figura 4.2.1: **Pontos de contato e pontos encobertos**. Dois objetos são observados por duas câmeras com centros de projeção $\mathbf{C}^p$ e $\mathbf{C}^q$. Suas posições são os pontos de contato $\mathbf{X}$ e $\mathbf{Y}$. O ponto $\mathbf{W}$ é encoberto pelo objeto em $\mathbf{Y}$ em relação à câmera q, o que lhe garante algum suporte. O ponto $\mathbf{Z}$ encontra-se totalmente encoberto, não sendo visto diretamente por nenhuma das câmeras. Este ponto obtém suporte dos dois objetos e pode acumular tanto suporte quanto um ponto de contato.

Algoritmo 2 Cômputo de suporte utilizando a Transformada de Hough com Q câmeras

1: **procedimento** THS($C_q = \langle \mathbf{v}_Z^q, \mathtt{H}^q, F^q \rangle$, para $q = 1..Q$, M, N)
2: $A \leftarrow$ NovoAcumulador(M, N) ▷ M e N definem a resolução do acumulador
3: **para** $q \leftarrow 1, Q$ **faça**
4: $\mathbf{V}_Z^q \leftarrow \mathtt{H}^q \mathbf{v}_Z^q$
5: **para** $\mathbf{x} \in F^q$ **faça**
6: $\mathbf{X} \leftarrow \mathtt{H}^q \mathbf{x}$
7: **para** $\mathbf{X}_i \in \langle \mathbf{X}, \mathbf{V}_Z^q \rangle$ **faça**
8: $A[\mathbf{X}_i] \leftarrow A[\mathbf{X}_i] + 1$
9: **devolva** A

aos 3 indivíduos em cena. Porém, alguns problemas podem ser notados.

A retas definidas pela relação formam feixes que se intersectam em $\mathbf{V}_Z^q$, a projeção do ponto afim $\mathbf{v}_Z^q$ em Π. A concentração de retas nos acumuladores próximos a $\mathbf{V}_Z^q$ faz com que essas células acumulem votos de muitos pixeis, como pode ser visto, por exemplo, no canto superior direito da Figura 4.2.2.

Pontos encobertos também produzem máximos locais. Isso não representaria um problema em regiões vistas por muitas câmeras. Porém, havendo poucas câmeras isto pode representar uma fonte considerável de falsos-positivos. No exemplo da Figura 4.2.2, só a região central da planta é vista pelas 3 câmeras. Cerca de 2/3 da área de interesse é visível por apenas 2 câmeras de forma que, havendo vários objetos em cena, é comum que várias regiões encontrem-se encobertas simultaneamente nas 2 imagens, como foi ilustrado na Figura 4.2.1.

4.3 Extensões do algoritmo básico

O Algoritmo 2 pode ser progressivamente estendido a medida que mais informação sobre a cena é conhecida. As informações são utilizadas para restringir o mapeamento, produzindo concentração de evidência em locais mais verossímeis.

4.3.1 Orientação do plano de imagem

Como discutido na Seção 4.1.1, uma das informações mais básicas sobre as câmeras é a orientação da imagem, ou seja, o sentido do vetor normal ao plano Π. Com essa informação, a reta do mapeamento imposto na Equação 4.14 pode ser restrita a uma semi-reta. Em termos físicos, define-se o local mais distante da câmera no qual um objeto pode ser colocado de forma que ainda seja capaz de encobrir o pixel $\mathbf{x}$: o ponto $\mathbf{X} = \mathtt{H}^q\mathbf{x}$.

Havendo uma orientação u^q definida para cada câmera, o Algoritmo 3 pode ser utilizado. A Figura 4.3.1 exibe o resultado obtido. Comparado ao resultado do Algoritmo 2, é possível notar que não há mais acúmulo de suporte em regiões próximas aos pontos $\mathbf{V}_Z^q$.

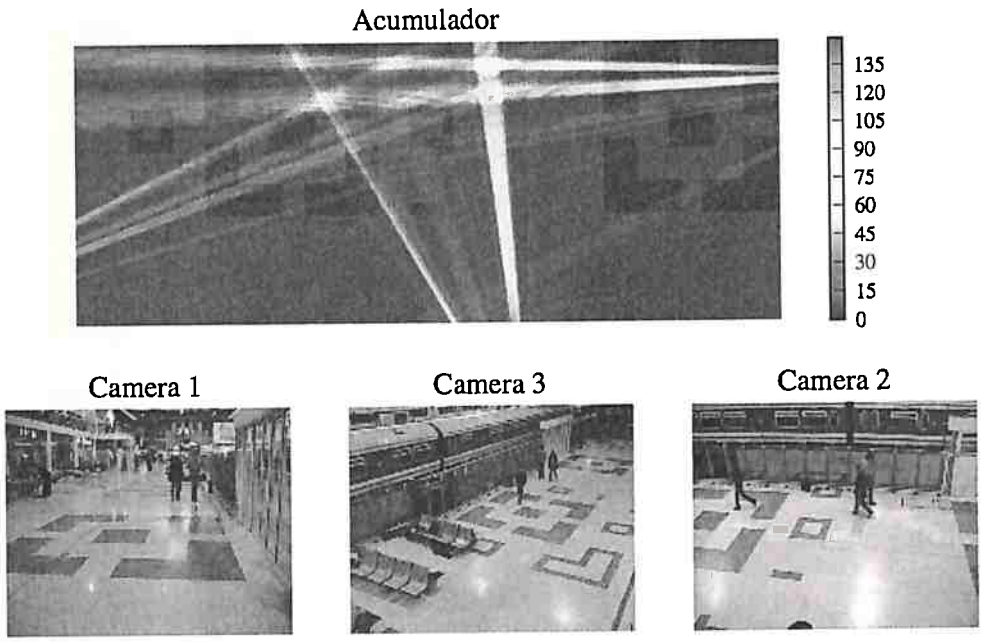


Figura 4.2.2: **Resultado obtido pelo Algoritmo 2.** Quadro 2562 da sequência S07 (PETS 2006). Há três pessoas na região de interesse, vistas em oclusão total ou parcial em duas das três câmeras. O acumulador A devolvido pelo Algoritmo 2 é exibido na parte superior da figura. É possível identificar máximos locais nas posições do 3 indivíduos. Porém, outros máximos locais podem ser vistos em regiões encobertas e nos limites da grade de acumuladores (regiões próximas a $\mathbf{V}_Z^q$).

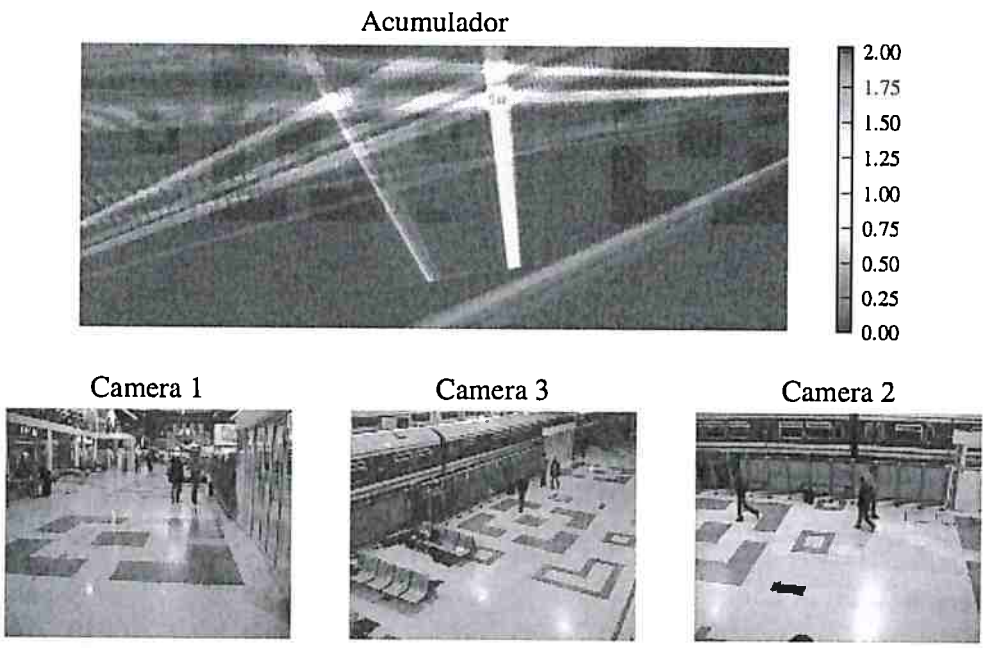


Figura 4.3.1: **Resultado obtido restringindo-se o sentido do mapeamento.** Quadro 2562 da sequência S07 (PETS 2006), visto anteriormente na Figura 4.2.2. O acumulador A obtido é exibido na parte superior da figura, onde o suporte do ponto é computado considerando-se apenas os pixeis acima dele (vide texto para detalhes). Por consequência, não há mais acúmulo de suporte próximo às projeções dos pontos afim, $\mathbf{V}_Z^q$.

Algoritmo 3 Cômputo de suporte utilizando a Transformada de Hough com Q câmeras. A orientação u^q de cada plano de imagem é conhecida

procedimento THS2($C_q = \langle \mathbf{v}_Z^q, \mathtt{H}^q, F^q \rangle$, para $q = 1..Q$, M, N, u)
  $A \leftarrow$ NovoAcumulador(M, N) ▷ M e N definem a resolução do acumulador
  para $q \leftarrow 1, Q$ **faça**
    para $\mathbf{x} \in F^q$ **faça**
      $\mathbf{X} \leftarrow \mathtt{H}^q\mathbf{x}$
      $\vec{\mathbf{N}}_\mathbf{x} \leftarrow \mathtt{H}^q\vec{\mathbf{n}}_\mathbf{x}$
      para $\mathbf{X}_i \in [\mathbf{X}, u^q\vec{\mathbf{N}}_\mathbf{x}\rangle$ **faça** ▷ Semi-reta com origem em $\mathbf{X}$
        $A[\mathbf{X}_i] \leftarrow A[\mathbf{X}_i] + 1$
  devolva A

4.3.2 Altura máxima de um objeto visível

A posição mais afastada da câmera na qual um objeto pode conter $\mathbf{x}$ é o próprio ponto em si, ou seja, $\mathbf{x}_\perp = \mathbf{x}$. Uma questão que surge naturalmente é qual o ponto mais *próximo* da câmera que poderia conter $\mathbf{x}$?

Como visto na Seção 4.1.3, a localização mais próxima possível à câmera ocorre quando $\mathbf{x}$ corresponde ao topo do objeto ("cabeça") e o objeto apresenta a maior altura observável, $h_{\max}$. O comprimento que corresponde a $h_{\max}$, em pixeis, para cada posição, é dado por $\mathfrak{h}_{\max}(\cdot)$, de modo que a localização procurada é dada por $\mathbf{x}_\perp = \mathbf{x} - \vec{\mathbf{n}}_\mathbf{x}\mathfrak{h}_{\max}(\mathbf{x})$. Desta forma, é possível limitar o número de localizações visitadas na grade de acumuladores a um intervalo mais restrito. Essa é a estratégia utilizada pelo Algoritmo 4.

Comparando-se o resultado obtido por esse algoritmo, exibido na Figura 4.3.2, aos resultados vistos previamente, é possível notar o efeito drástico causado nos valores observados nos acumuladores. Para cada pixel de figura, o número de pontos visitados no espaço paramétrico é muito mais limitado, produzindo ganhos em eficiência (menos células acumuladoras são visitadas) e na acurácia da localização.

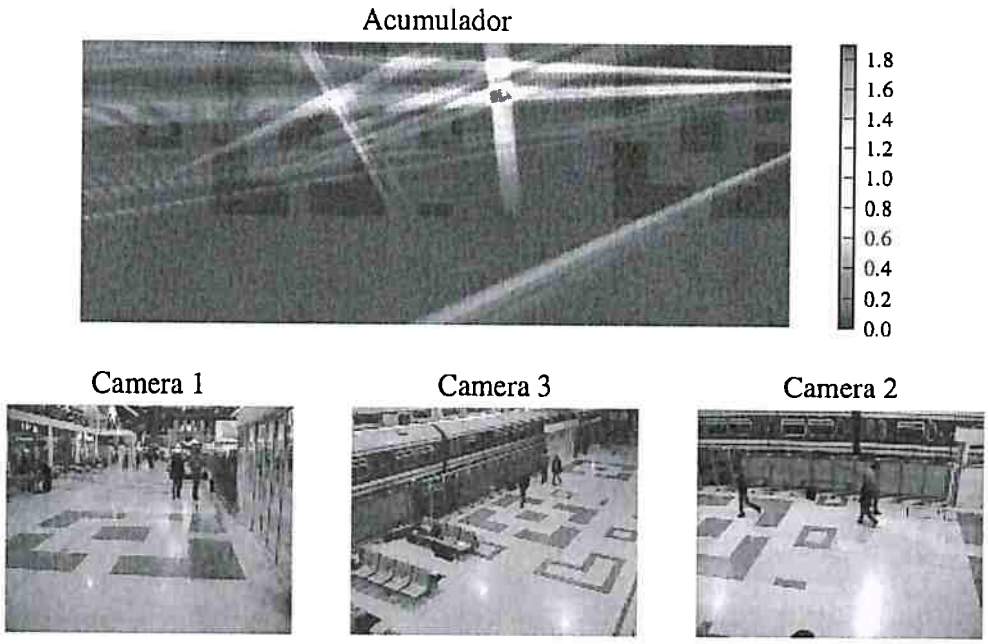


Figura 4.3.2: **Resultado obtido pelo Algoritmo 4.** Quadro 2562 da sequência S07 (PETS 2006). O acumulador A devolvido pelo Algoritmo 4 é exibido na parte superior da figura. É possível identificar com maior clareza máximos locais nas posições do 3 indivíduos. Durante a transformada, o algoritmo é capaz de definir a posição mais próxima à câmera e a mais afastada onde um objeto qualquer seria capaz de encobrir um determinado pixel.

Algoritmo 4 Algoritmo extendido para a Transformada de Hough com Q câmeras

procedimento THSX($\mathcal{C}_q = \langle \mathbf{v}_Z^q, \mathtt{H}^q, F^q \rangle$, para $q = 1..Q$, M, N)
  $A \leftarrow$ NovoAcumulador(M, N)
  para $q \leftarrow 1, Q$ **faça**
    para $\mathbf{x} \in F^q$ **faça**
      $\mathbf{x}_\perp \leftarrow \mathbf{x} - \mathfrak{h}_{\max}(\mathbf{x}) \cdot \vec{\mathbf{n}}_\mathbf{x}$
      $\mathbf{X}_1 \leftarrow \mathtt{H}^q \mathbf{x}_\perp$ ▷ Posição mais próxima da câmera onde um objeto pode encobrir $\mathbf{x}$
      $\mathbf{X}_2 \leftarrow \mathtt{H}^q \mathbf{x}$ ▷ Posição mais afastada da câmera onde um objeto pode encobrir $\mathbf{x}$
      para $\mathbf{X}_i \in \langle \mathbf{X}_1, \mathbf{X}_2 \rangle$ **faça**
        $A[\mathbf{X}_i] \leftarrow A[\mathbf{X}_i] + 1$
  devolva A

4.3.3 Descontinuidades

Os algoritmos vistos até aqui não levam em consideração descontinuidades existentes nas figuras observadas. Considere novamente $\mathbf{x}$ um ponto de figura e $\mathbf{X}$ é uma posição em Π onde um objeto seria capaz de encobrir $\mathbf{x}$ no plano de imagem. O ponto $\mathbf{x}$ contribui com o acumulador de $\mathbf{X}$ s independentemente do espaço vazio (fundo) que possa existir entre $\mathbf{x}_\perp = \mathtt{H}^{-1}\mathbf{X}$ e $\mathbf{x}$. Em outras palavras, assume-se que um objeto possa ser descontínuo.

Embora esta característica seja útil quando ocorre camuflagem (Capítulo 3) ou outros falsos-negativos na subtração de fundo, ela é fisicamente contra-intuitiva dado que os objetos são conexos. Essa é a característica responsável pelas longas "caudas" vistas ao redor dos máximos locais nas Figuras 4.2.2, 4.3.1 e 4.3.2.

Para penalizar descontinuidades, a razão entre pixeis de *foreground* e o total de pixeis no segmento entre $\mathbf{x}$ e $\mathbf{x}_\perp$ pode ser computada. Seja

$$\mathfrak{f}_\mathbf{x}(\mathbf{X}) = \frac{\text{número de pixeis de figura em } [\mathbf{x}, \mathbf{x}_\perp]}{\text{número total de pixeis em } [\mathbf{x}, \mathbf{x}_\perp]}. \tag{4.15}$$

O Algoritmo 4.3.2 pode ser então alterado de modo a penalizar as descontinuidades

observadas ao se assumir que um objeto em $\mathbf{X}_i$ encobre $\mathbf{x}$. Deriva-se daí o Algoritmo 5.

Algoritmo 5 Algoritmo extendido para a Transformada de Hough com Q câmeras

procedimento THSX($C_q = \langle \mathbf{v}_Z^q, \mathtt{H}^q, F^q \rangle$, para $q = 1..Q$, M, N)
  $A \leftarrow$ NovoAcumulador(M, N)
  para $q \leftarrow 1, Q$ **faça**
    para $\mathbf{x} \in F^q$ **faça**
      $\mathbf{x}_\perp \leftarrow \mathbf{x} - \mathfrak{h}_{\max}(\mathbf{x}) \cdot \vec{\mathbf{n}}_\mathbf{x}$
      $\mathbf{X}_1 \leftarrow \mathtt{H}^q \mathbf{x}_\perp$ ▷ Posição mais próxima da câmera onde um objeto pode encobrir $\mathbf{x}$
      $\mathbf{X}_2 \leftarrow \mathtt{H}^q \mathbf{x}$ ▷ Posição mais afastada da câmera onde um objeto pode encobrir $\mathbf{x}$
      para $\mathbf{X}_i \in \langle \mathbf{X}_1, \mathbf{X}_2 \rangle$ **faça**
        $A[\mathbf{X}_i] \leftarrow A[\mathbf{X}_i] + \mathfrak{f}_\mathbf{x}(\mathbf{X}_i)$
  devolva A

A Figura 4.3.3 exibe o resultado obtido pela modificação. É possível ver claramente, comparado ao resultado da Figura 4.3.2, a redução das "caudas" e o maior contraste entre os máximos locais nos pontos de contato e o restante do acumulador.

4.4 Obtenção de máximos locais por deslocamento à média

Uma vez computados os valores na grade de acumuladores, os máximos locais devem ser localizados. Primeiramente, uma limiarização é aplicada, suprimindo qualquer ponto $\mathbf{X}$ que apresente suporte inferior a $\mathfrak{v}(\mathbf{X})h_{\min}$, garantindo que os objetos apresentem uma altura mínima. Então, os máximos locais são identificados pelo procedimento de deslocamento à média.

Deslocamento à média (*mean-shift*) é um procedimento simples que move cada ponto de

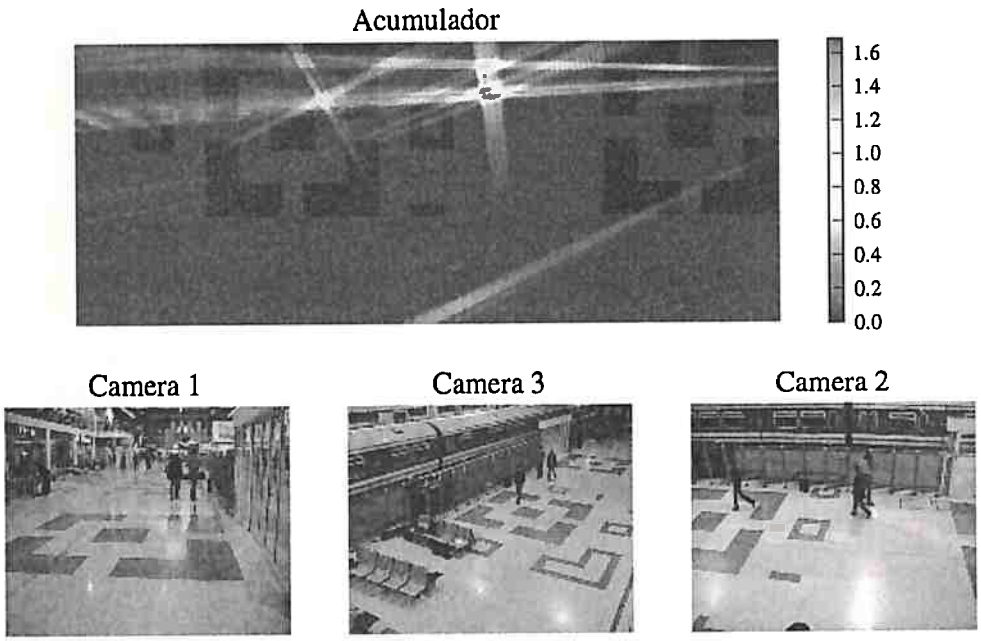


Figura 4.3.3: **Penalização das descontinuidades na figura.** Assumindo que o módulo de subtração de fundo não produz falsos-negativos, a figura observada para os objetos deveria ser contínua. Incorporando a razão entre pixeis de figura e pixeis observados (Equação 4.15) ao cálculo do suporte, as descontinuidade são penalizadas. Isso reduz as "caudas" observadas ao redor das posições dos indivíduos.

um conjunto de dados para a média ponderada dos pontos em sua vizinhança [8]. A técnica foi introduzida por Fukunaga e Hostetler [20] como método para estimação do gradiente de uma função de densidade. Ela foi re-introduzida por Comaniciu e Meer [9, 10], quando ganhou a atenção da comunidade de visão computacional, apesar de Cheng [8] ter anteriormente descrito suas aplicações em *clustering*, otimização e na própria Transformada de Hough.

Seja $\mathcal{X}$ o conjunto de pontos $\mathbf{X}_i$ tais que $A[\mathbf{X}_i] \geq k_a$, onde k_a é um limiar (baseado na altura mínima, por exemplo). O acumulador A pode ser visto como uma *função de peso* $\mathcal{X}$. A média obtida com um núcleo de convolução K no ponto $\mathbf{X} \in \mathcal{X}$ é definida como

$$m(\mathbf{X}) = \frac{\sum_{\mathbf{Y} \in \mathcal{X}} K(\mathbf{Y} - \mathbf{X}) A[\mathbf{Y}] \mathbf{Y}}{\sum_{\mathbf{Y} \in \mathcal{X}} K(\mathbf{Y} - \mathbf{X}) A[\mathbf{Y}]}. \tag{4.16}$$

Considere $\mathcal{T}$ um conjunto de "candidatos" (inicialmente, $\mathcal{T} = \mathcal{X}$). O algoritmo de deslocamento à média consiste em uma sequência de iterações da forma $\mathcal{T} \leftarrow m(\mathcal{T})$, sendo $m(\mathcal{T}) = \{m(\mathbf{T}); \mathbf{T} \in \mathcal{T}\}$. Idealmente, o algoritmo termina quando $m(\mathcal{T}) = \mathcal{T}$. Para evitar instabilidade numérica, o procedimento é finalizado quando $\|m(\mathbf{T}) - \mathbf{T}\| \leq \epsilon$ para algum limiar ϵ escolhido.

O que torna o deslocamento à média útil no presente contexto é que ele é um capaz de encontrar todos os máximos locais existentes em $\sum_{\mathbf{Y} \in \mathcal{X}} H(\mathbf{Y} - \mathbf{X}) A[\mathbf{Y}]$, onde H é um núcleo de convolução *sombra* (*shadow*) de K. Se o núcleo de convolução K for Gaussiano, $H = K$ e o procedimento irá encontrar todos os máximos locais em *um versão suavizada de* A: a convolução de A por uma função Gaussiana.

Uma análise detalhada sobre as propriedades do deslocamento a média, incluindo provas sobre sua capacidade de localização de máximos e sua convergência, pode ser encontrada no trabalho de Cheng [8]. A Figura 4.4.1 exibe o resultado obtido na localização dos máximos locais para o exemplo corrente. Após 3 iterações do procedimento, o algoritmo converge e as posições dos três indivíduos são encontradas com sucesso.

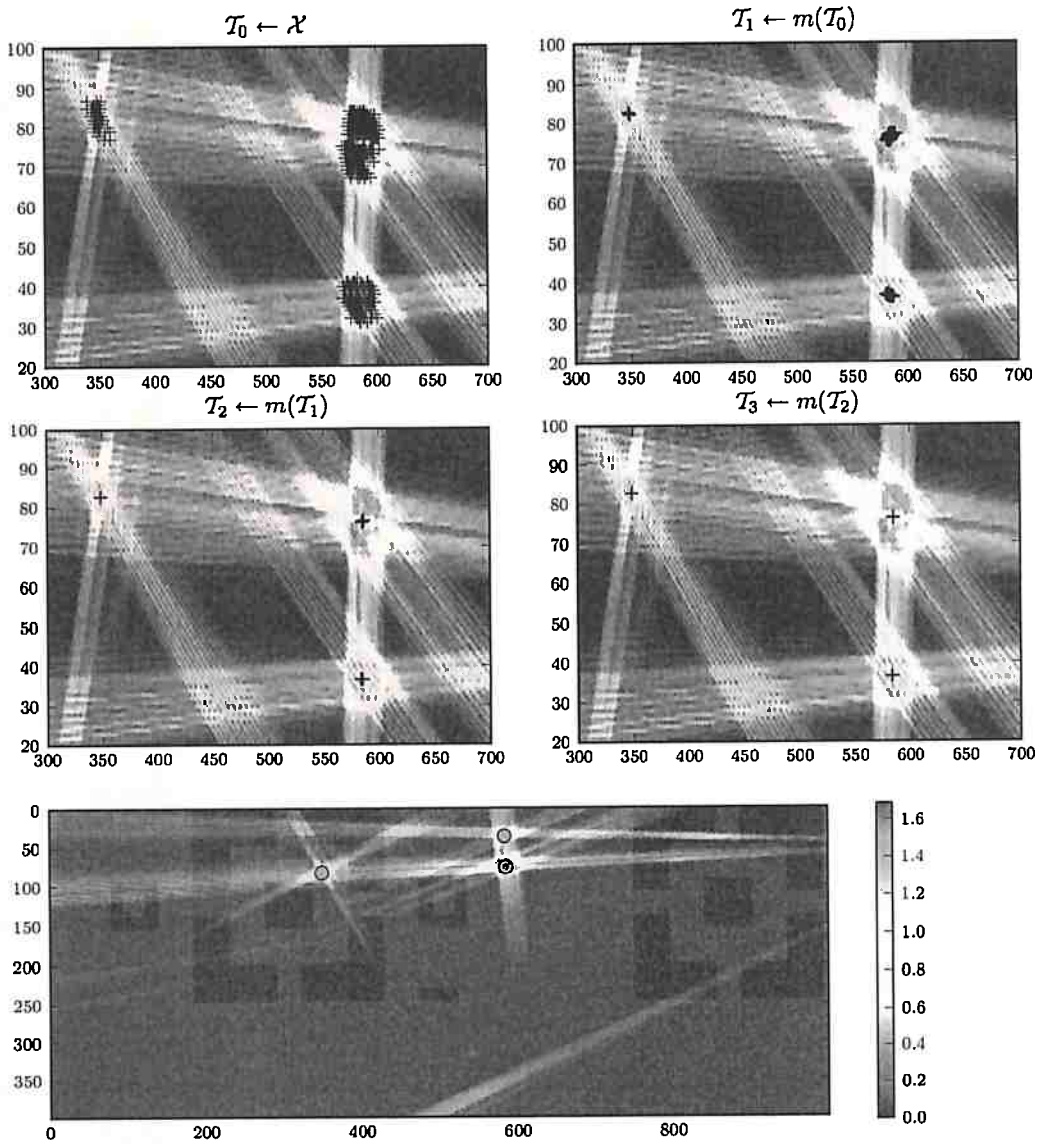


Figura 4.4.1: **Deslocamento à média para localização de máximos locais.** O acumulador exibido anteriormente na Figura 4.3.3 é utilizado como função peso. O conjunto inicial é obtido por limiarização: $\mathcal{X} = \{\mathbf{X}; A[\mathbf{X}] > 1.2\}$. O núcleo de convolução utilizado é um *kernel Gaussiano truncado*: $G^{\beta}F\lambda(\mathbf{X}) = e^{-\beta\|X\|^2}$ se $\|X\| \leq \lambda$, 0 caso contrário. No exemplo, $\beta = 0.05$ e $\lambda = 25$. O algoritmo converge após 3 iterações e os resultados finais, exibidos na parte inferior da figura, coincidem com as posições dos três indivíduos vistos em cena.

4.5 Localização de pessoas com imagens retificadas

A retificação de imagens [25] é um processo de transformação utilizado para alinhar múltiplas imagens. Esse processo é utilizado em visão computacional estéreo para tornar linhas epipolares paralelas ao eixo horizontal da imagem e dessa forma facilitar o cálculo de correspondências. Retificação de imagens é também largamente utilizado em Sistemas de Informação Geográfica (GIS) para a composição de múltiplas imagens em um único sistema comum de coordenadas (um mapa).

Assim como em visão estéreo, essa seção descreve como a utilização de imagens retificadas no cálculo de suporte torna o processo de localização de pessoas mais eficaz.

4.5.1 Retificação de imagens

O objetivo aqui é retificar o plano de imagem na direção normal ao plano Π, representada pelo ponto ideal $\mathbf{v}_Z$, minimizando distorções para que a imagem retificada mantenha tantas propriedades da imagem original quanto possível. Como discutido no Capítulo 3, um plano de imagem ortogonal a Π apresentaria o ponto $\mathbf{v}_Z$ no infinito, e um único vetor representaria a direção normal para todo o plano de imagem, ou seja, retas ortogonais a Π seriam imageadas como retas paralelas.

Em [25], Hartley desenvolve um método para retificação de pares de imagens estéreo que minimiza a distorção entre as imagens originais e as de entrada. Ele mostra que uma homografia H que leva um ponto da imagem a um ponto afim possui 4 graus de liberdade, e que é possível então restringir H de forma a minimizar as distorções perspectivas. No caso estéreo, as imagens devem ser retificadas para que a linha epipolar se torne horizontal. No entanto, para o caso de suporte, desejamos que a direção vertical seja retificada.

Afim de minimizar as distorções perspectivas, Hartley [25] impõe que H deve ser tão rígida quanto possível em torno de algum ponto de referência X_0. Uma escolha adequada para o cálculo de suporte seria o centro da imagem, de forma que a imagem retificada seja parecida com a imagem original da câmera.

a b

Figura 4.5.1: **Exemplo de imagem retificada.** (a) Imagem onde o ponto ideal $\mathbf{v}_Z$ é um ponto finito. A retificação é obtida através de uma transformação projetiva que mapeie $\mathbf{v}_Z$ para um ponto no infinito $(0, f, 0)^\top$. (b) Imagem obtida após a transformação. Ela equivale a um plano de imagem ortogonal ao plano do solo Π. Note como as bordas das janelas se tornam paralelas ao eixo das ordenadas do plano de imagem.

Assumindo que $\mathbf{v}_Z$ esteja em $(0, f, 1)^\top$, é fácil verificar que a seguinte transformação projetiva $\mathtt{H}_p$

$$\mathtt{H}_p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/f & 1 \end{bmatrix}. \tag{4.17}$$

leva $\mathbf{v}_Z$ ao ponto $(0, 1, 0)^\top$ no infinito, como desejado. Pode-se demonstrar também que essa transformação é quase rígida em torno de X_0 [25].

Para um ponto $\mathbf{v}_Z$ arbitrário, uma rotação em torno de X_0 é suficiente para colocar $\mathbf{v}_Z$ no formato $(0, f, 1)^\top$. Essa rotação, caracterizada pela transformação afim $\mathtt{H}_a$ pode ser combinada com $\mathtt{H}_p$ para criar a homografia desejada. Assim, a retificação pode ser obtida através da transformação projetiva dada pela matriz $\mathtt{H}_r = \mathtt{H}_p\mathtt{H}_a$. A Figura 4.5.1 ilustra o resultado obtido.

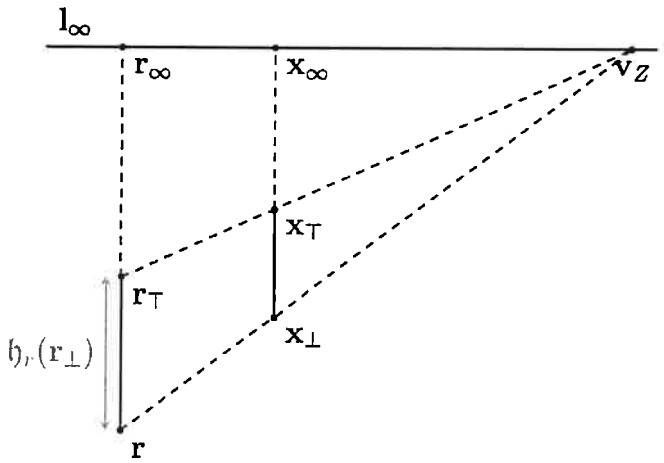


Figura 4.5.2: **Cálculo da altura em imagens retificadas**. Se a direção vertical ao solo foi retificada, então as razões entre os comprimentos nessa direção foram preservadas. Assim, definida a altura em pixeis $h_r(\mathbf{r}_\perp)$ de um objeto em uma posição de referência $\mathbf{r}_\perp$, a altura em pixeis $h_r(\mathbf{x}_\perp)$ do mesmo objeto, quando posicionado em $\mathbf{x}_\perp$ pode ser inferida por simples semelhança de triângulos.

4.5.2 Correção perspectiva

A retificação de uma câmera q como descrito em 4.5.1 leva o ponto afim na direção vertical para o infinito, fazendo com que todas as linhas verticais da cena apareçam paralelas ao eixo das ordenadas no plano de imagem de cada câmera q. Portanto, nas imagens retificadas, propriedades invariantes em transformações afins, como a razão de distâncias, são preservadas.

Como descrito na Seção 4.1.2, a correção perspectiva é necessária para acumular uma evidência que seja proporcional a altura do objeto. Dessa forma, objetos próximos, que ocupam uma grande região na imagem, recebem o mesmo peso que objetos afastados da câmera, que ocupam uma pequena região na imagem devido a distorção perspectiva. Na Seção 4.1.2, sem a retificação de imagens, foi utilizada a propriedade de razão cruzada para calcular a altura de objetos.

O uso de imagens retificadas permite que a correção perspectiva seja calculada de forma

mais eficiente, como a razão de distâncias, que é invariante no caso de transformações afins. A Figura 4.5.2 ilustra o processo. Na imagem retificada, a linha do horizonte de Π é transformada para a linha $\mathbf{l}_\infty$. Seja $\mathbf{r}_\perp = (x_r, y_r)$ a posição do objeto de referência, $\mathbf{r}_\infty = (x_{r_\infty}, y_{r_\infty})$ a projeção de $\mathbf{r}_\perp$ sobre $\mathbf{l}_\infty$ e h_r a altura do objeto. A altura *em pixeis* na posição $\mathbf{r}_r$ correspondente a h_r é conhecida: $\mathfrak{h}_r(\mathbf{r}_\perp)$. Considere $\mathbf{x}_\perp = (x, y)$ uma outra posição qualquer e $\mathbf{x}_\infty = (x_\infty, y_\infty)$ sua projeção na linha do horizonte. Então, a altura $\mathfrak{h}_r(\mathbf{x}_\perp)$ em pixeis do objeto de referência em $\mathbf{x}_\perp$ pode ser calculada por semelhança de triângulos como

$$\mathfrak{h}_r(\mathbf{x}_\perp) = \frac{\mathfrak{h}_r(\mathbf{r}_\perp)(y - y_\infty)}{y_r - y_{r_\infty}}, \tag{4.18}$$

dispensando o uso da razão cruzada. Com a preservação da razão entre distâncias, é possível estabelecer agora, para cada pixel $\mathbf{x}_\perp$, o valor $w(\mathbf{x}_\perp)$, correspondente ao comprimento de *um único pixel*, em relação a altura de referência h_r, isto é, $w(\mathbf{x}_\perp) = 1/\mathfrak{h}_r(\mathbf{x}_\perp)$. O valor de $w(\mathbf{x}_\perp)$ pode ser empregado no processo de acúmulo de evidência.

Retomando o visto no Algoritmo 2, o suporte no ponto $\mathbf{x}$ pode ser computado diretamente pela varredura das colunas no plano de imagem, sendo normalizado com o uso de $\mathfrak{h}_{\max}(\cdot)$. Obtêm-se assim o Algoritmo 6.

4.6 Filtragem de falsos-positivos

Como comentado anteriormente na Seção 4.2.2, pontos encobertos podem ser uma fonte de falsos-positivos, principalmente havendo poucas câmeras disponíveis. Contudo, como a posição de cada câmera q em Π é conhecida, dada por $\mathbf{V}_Z^q = \mathtt{H}^q \mathbf{v}_Z^q$ e havendo uma estimação da altura dos objetos, é possível definir relações de oclusão entre os objetos.

Considere um grafo $\mathcal{G}$ cujo conjunto de vértices V é formado pelos pontos encontrados pela Transformada de Hough (os máximos locais de suporte). Há um conjunto de arestas E^q definido para cada câmera. A aresta $(\mathbf{X}, \mathbf{Y})$ se encontra em E^q se e somente se um objeto de altura $h_\mathbf{X} = A[\mathbf{X}]/\mathfrak{v}(\mathbf{X})$ posicionado em $\mathbf{X}$ for capaz de encobrir um objeto de altura $h_\mathbf{Y} = A[\mathbf{Y}]/\mathfrak{v}(\mathbf{Y})$ posicionado em $\mathbf{Y}$.

Algoritmo 6 Algoritmo para cômputo do suporte acumulado através de imagens retificadas.

procedimento SUPORTE(F, $\mathfrak{h}_{\max}$, $\mathfrak{h}_{\min}$, m, n)
  para $x \leftarrow 1, n$ **faça**
    $C[x, 0] \leftarrow 0$
    para $y \leftarrow 1, m$ **faça**
      se $(x, y) \in F$ **então**
        $C[x, y] \leftarrow C[x, y-1] + 1$
      senão
        $C[x, y] \leftarrow C[x, y-1]$
      se $C[x, y] < \mathfrak{h}_{\min}(x, y)$ **então**
        $S[x, y] \leftarrow 0$
      senão
        se $C[x, y] > \mathfrak{h}_{\max}(x, y)$ **então**
          $S[x, y] \leftarrow 1$
        senão
          $S[x, y] \leftarrow C[x, y] / \mathfrak{h}_{\max}(x, y)$
  devolva S
procedimento INTEGRAÇÃO($\{C_q\}_{q=1..Q}$, M, N)
  $A \leftarrow$ NOVOACUMULADOR(M, N)
  para $q \leftarrow 1, Q$ **faça**
    $S^q \leftarrow$ SUPORTE(F^q, $\mathfrak{h}^q_{\max}$, $\mathfrak{h}^q_{\min}$, m^q, n^q)
  para $\mathbf{X} \in M \times N$ **faça**
    $A[\mathbf{X}] \leftarrow \sum_{q=1}^{Q} S^q[\mathtt{H}^{-1}\mathbf{X}]$
  devolva A

A partir do grafo $\mathcal{G}$, o Algoritmo 7 classifica cada ponto $\mathbf{X}$ como um ponto de contato ou como um ponto encoberto. Inicialmente, a classificação de um ponto é desconhecida. A heurística utilizada é tomar, a cada iteração, o ponto ainda não classificado que apresenta o maior suporte. Tal ponto é classificado como um ponto de contato e as relações de cobertura são re-avaliadas. Se um ponto desconhecido se encontra encoberto por algum ponto de contato em todas as câmeras, ele será classificado como encoberto. O procedimento é repetido até que todos os pontos sejam classificados.

Algoritmo 7 Algoritmo para filtragem de pontos encobertos

```
1:  procedimento COBERTURA(G = ⟨V, E^q⟩, q = 1..Q)
2:      para X ∈ V faça
3:          estado[X] ← DESCONHECIDO
4:          INSIRA(Q, X, A)
5:      enquanto Q ≠ ∅ faça
6:          X ← POP(Q)
7:          estado[X] ← CONTATO
8:          para q ← 1..Q faça
9:              para (X, Y) ∈ E^q faça
10:                 se estado[Y] = DESCONHECIDO então
11:                     se Y ocluso por algum ponto CONTATO em todas as câmeras então
12:                         estado[X] ← ENCOBERTO
13:         REMOVECONHECIDOS(Q)
14:     devolva {X : estado[X] = CONTATO}
```

Vários métodos foram propostos no presente capítulo para a detecção de objetos, mesmo em oclusão total ou parcial, através de múltiplas câmeras. Utilizou-se como entrada os resultados de um subtrator de fundo, o ponto afim que define a direção vertical e homografias entre planos. No próximo capítulo, o sistema será completado por um módulo de rastreamento que combina um modelo de aparência multi-câmera a Filtros de Kalman. Tal módulo

é capaz de empregar as localizações detectadas e, se disponível, a altura recuperada na a construção do modelo de aparência de cada objeto rastreado.

Capítulo 5

Rastreamento de múltiplos objetos

Um problema inerente ao rastreamento de múltiplos objetos é a associação das observações detectadas e às trajetórias em curso [4, 55]. No caso de filtros de Kalman e filtros de partículas, uma das principais questões é definir qual observação escolher para a atualização de cada filtro, baseado no estado previsto pelos mesmos. O problema é ilustrado na Figura 5.0.1, retirada do artigo de Blackman [4], na qual três observações **X**, **Y** e **Z** devem ser associadas a dois filtros, que produziram as predições $\mathbf{T}_1$ e $\mathbf{T}_2$ para o instante em questão.

O sistema proposto rastreia vários indivíduos simultaneamente empregando um filtro de Kalman para cada pessoa. Cada indivíduo é representado por um modelo de aparência para cada câmera. Cada modelo consiste de histogramas que representam a distribuição de cor encontrada em duas partições do corpo do indivíduo, torso e pernas. Cada modelo também mantém máscaras identificando os pixeis visíveis do objeto. Os histogramas e as máscaras são atualizadas quadro a quadro. Os modelos de aparência e as predições do filtros de Kalman são utilizadas em um algoritmo guloso onde cada observação é associada a um único filtro, que tem seu estado e modelos atualizados.

5.1 Filtros de Kalman

O filtro de Kalman fornece uma solução recursiva eficiente para a estimação do estado de um processo. Uma introdução geral aos filtros de Kalman pode ser encontrada nos trabalhos de

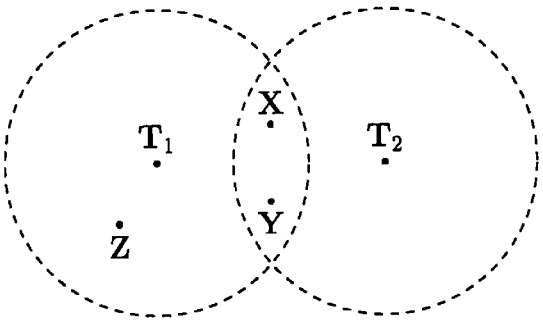


Figura 5.0.1: **Situação de conflito durante associação**. Três observações, **X**, **Y** e **Z** são obtidas durante a etapa de detecção. Dois rastreadores estão sendo avaliados, fornecendo as predições $\mathbf{T}_1$ e $\mathbf{T}_2$. Todas as observações estão dentro da área de associação do filtro responsável pela predição $\mathbf{T}_1$. As observações **X** e **Y** se encontram também na área de associação do filtro responsável por $\mathbf{T}_2$. O problema consiste em associar uma única observação a cada filtro. Esta figura é similar à exibida por Blackman [4].

Maybeck [36] e Welch e Bishop [52] enquanto que sua aplicação ao rastreamento de objetos é abordada com maior profundidade no trabalho de Bar-Shalom *et al.* [2].

Na solução proposta nesse trabalho, o *estado* de um objeto no instante t é representado pelo vetor

$$\mathbf{s}_t = (X, Y, \Delta X, \Delta Y)^\mathsf{T} \tag{5.1}$$

em que (X, Y) representa a posição do objeto no plano do solo Π e $(\Delta X, \Delta Y)$ corresponde a velocidade instantânea do objeto. O processo sendo estimado é definido pelo *modelo de dinâmica*

$$\mathbf{s}_t = \mathtt{A}\mathbf{s}_{t-1} + \mathbf{w}_{t-1} \tag{5.2}$$

onde $\mathtt{A}$ é a matriz que governa a dinâmica do objeto, aqui definida por um modelo de velocidade constante por

$$\mathtt{A} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{5.3}$$

enquanto que $\mathbf{w}_t$ é o *ruído do processo*, que obedece a uma distribuição normal de média zero

$$p(\mathbf{w}_t) \sim N(0, \mathtt{Q}), \tag{5.4}$$

onde $\mathtt{Q}$ é a matriz de covariância do ruído do processo. Em outras palavras, no instante seguinte espera-se que o objeto se encontre em $(X + \Delta X, Y + \Delta Y)$, a menos de um ruído Gaussiano, seguindo assim um modelo de movimento uniforme.

O processo é estimado a partir de observações (*medições*) que aqui são as posições $\mathbf{X}_t$ obtidas pelo método de detecção apresentado no Capítulo 4, em coordenadas não homogêneas, $\mathbf{X}_t = (X, Y)$. Essas observações se relacionam ao estado do objeto através do *modelo de observação*

$$\mathbf{X}_t = \mathtt{G}\mathbf{s}_t + \mathbf{W}_t \tag{5.5}$$

onde $\mathtt{G}$ é dada por

$$\mathtt{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{5.6}$$

enquanto que $\mathbf{W}_t$ é o ruído de medição governado pela distribuição normal com média zero

$$p(\mathbf{W}_t) \sim N(0, \mathtt{R}), \tag{5.7}$$

em que $\mathtt{R}$ é a matriz de covariância do ruído de medição. Para modelar a covariância dos ruídos são utilizadas matrizes diagonais, nas quais os valores na diagonal principal são todos iguais a uma constante k_η.

Seja $\hat{\mathbf{s}}_t^-$ a estimação *a priori* (predição) do estado no instante t, obtida através da informação disponível até o mesmo, e $\hat{\mathbf{s}}_t$ a estimação *a posteriori*, obtida após a observação $\mathbf{X}_t$ ser considerada. Ambas as estimativas apresentam erros em relação ao estado real $\mathbf{s}_t$. A covariância do erro na predição é dada por

$$\mathtt{P}_t^- = E[(\mathbf{s}_t - \hat{\mathbf{s}}_t^-)(\mathbf{s}_t - \hat{\mathbf{s}}_t^-)^\top], \tag{5.8}$$

em que $E[x]$ corresponde à esperança de x. Similarmente, a covariância do erro na estimação *a posteriori* é dada por

$$\mathtt{P}_t = E[(\mathbf{s}_t - \hat{\mathbf{s}}_t)(\mathbf{s}_t - \hat{\mathbf{s}}_t)^\top]. \tag{5.9}$$

A estimação *a posteriori* é obtida através de uma combinação linear da predição e uma diferença ponderada entre a medida observada $\mathbf{X}_t$ e a medição prevista $\mathtt{G}\hat{\mathtt{s}}_t^-$:

$$\hat{\mathtt{s}}_t = \hat{\mathtt{s}}_t^- + \mathtt{K}_t(\mathbf{X}_t - \mathtt{G}\hat{\mathtt{s}}_t^-). \tag{5.10}$$

A matriz $\mathtt{K}_t$ que minimiza a covariância do erro *a posteriori* definida em (5.9) é dada por

$$\mathtt{K}_t = \mathtt{P}_t^-\mathtt{G}^\top(\mathtt{G}\mathtt{P}_t^-\mathtt{G}^\top + \mathtt{R})^{-1}. \tag{5.11}$$

O procedimento completo de predição e correção do filtro de Kalman é resumido na Figura 5.1.1. Assim, cada rastreador $\mathcal{T}_i$ possui armazenadas, no instante t, a estimativa para o estado e a matriz de covariância do erro no instante anterior, $\hat{\mathtt{s}}_{t-1}$ e $\mathtt{P}_{t-1}$. Para completarmos o procedimento, precisamos definir o método de associação, que atribui uma observação $\mathbf{X}_j$ e cada rastreador $\mathcal{T}_i$. Para tanto, precisamos definir o modelo de aparência utilizado.

5.2 Modelo de aparência

O modelo de aparência de um objeto visto por múltiplas câmeras $\mathcal{A} = \{M^q, H^q_{\text{torso}}, H^q_{\text{pernas}}\}_{q=1..Q}$ é composto por

- máscaras M^q que indicam quais são os pixeis que constituem o objeto modelado *e* que não estão oclusos na imagem da câmera q e
- histogramas de cor normalizados H^q_{torso} e H^q_{pernas} que representam a distribuição de cor observada nos pixeis visíveis, referentes às regiões do torso e das pernas do indivíduo.

O uso de histogramas normalizados visa produzir uma representação de cor que não se altere com o tamanho em pixeis do objeto representado. Como visto no Capítulo 3, objetos mais afastados da câmera apresentarão um número menor de pixeis. Para levar este efeito perspectiva em conta, o armazenamento e a manutenção de modelos de aparência baseados em janelas [23, 38, 44] teriam que se tornar mais complexos, o que justifica a escolha dos histogramas normalizados.

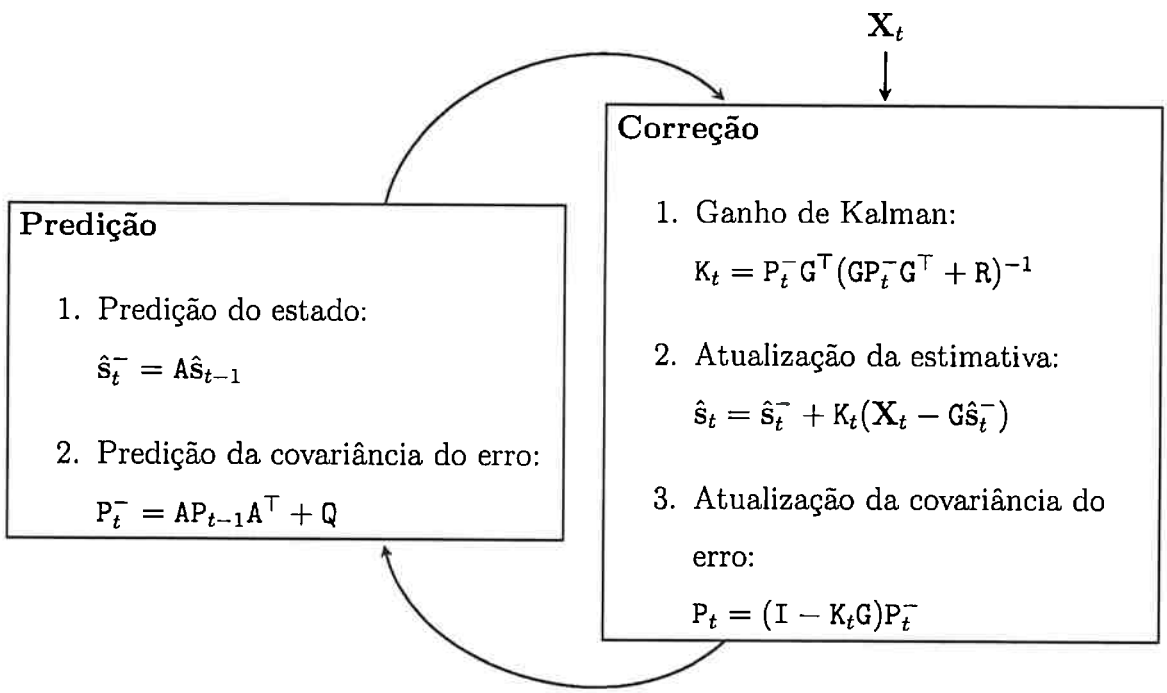


Figura 5.1.1: **Filtro de Kalman.** A figura mostra o procedimento para predição e correção de estados pelo filtro de Kalman (ver texto para detalhes). A figura é uma adaptação da exibida por Welch e Bishop [52].

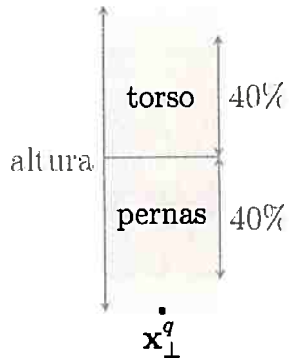


Figura 5.2.1: **Regiões do modelo de aparência: torso e pernas.**

5.2.1 Construção do modelo para uma observação em X

Considere que um objeto foi localizado na posição $\mathbf{X}$ que corresponde ao ponto $\mathbf{x}_{\perp}^{q}$ no plano de imagem da câmera q. A posição $\mathbf{x}_{\perp}^{q}$ e a altura do objeto definem uma janela retangular no plano de imagem[1]. As regiões correspondentes ao torso e às pernas são faixas com 40% da altura do objeto situadas acima e abaixo da "cintura" (o centro da janela), como ilustrado na Figura 5.2.1. Em execução, verificamos que esta formulação consegue marcar as regiões do torso e das pernas dos indivíduo, removendo as regiões correspondentes à cabeça e ao pés.

A utilização dessas duas regiões é vantajosa frente ao uso de um único histograma. Se um único histograma fosse empregado, um indivíduo de camisa branca e calças pretas apresentaria um modelo similar ao de um indivíduo de camisa preta e calças brancas, levando o sistema a possíveis erros de associação ao tratar duas pessoas nitidamente distintas ao olhar humano. A função da máscara M^q é definir os pixeis na faixa do torso e das pernas a serem considerados no cômputo dos histogramas.

5.2.2 Comparação de modelos

Os modelos H^q_{torso} e H^q_{pernas} são compostos por três histogramas h_c quantizados em 16 partições, uma para cada canal c em RGB. A *intersecção de histogramas* é utilizada na comparação,

[1] O eixo principal dessa janela retangular, contudo, não se encontra na direção do eixo das ordenadas no plano de imagem, mas é determinado pelo ponto no infinito $\mathbf{v}_Z$, como discutido no Capítulo 3.

definida por

$$\text{hsim}(h_i, h_j) = \sum_{b=1}^{16} \min(h_i[b], h_j[b]) \tag{5.12}$$

que produz valores entre 0 e 1 para histogramas normalizados (0 indica total dissimilaridade e 1 é produzido por histogramas idênticos).

Para avaliar os três canais, optou-se aqui pelo mínimo das intersecções, ou seja, a similaridade entre canais csim para dois modelos (H_i, H_j) relativos a uma mesma câmera q é dada pelo mínimo entre a similaridade na região do torso e a similaridade na região da pernas.

$$\text{csim}(H_i, H_j) = \min_{c=R,G,B} \{\text{hsim}(h_{c,i}, h_{c,j})\}. \tag{5.13}$$

Assim, se dois indivíduos utilizarem camisas de cores similares mas trajarem calças com cores distintas, o sistema será capaz de diferenciá-los e determinar associações mais corretas.

Finalmente, a similaridade entre dois modelos de aparência em múltiplas câmeras $\mathcal{A}_i$, $\mathcal{A}_j$ é dado por

$$\text{sim}(\mathcal{A}_i, \mathcal{A}_j) = \max_{q=1..Q} \{\min\{\text{csim}(H^q_{\text{torso},i}, H^q_{\text{torso},j}), \text{csim}(H^q_{\text{pernas},i}, H^q_{\text{pernas},j})\}\}. \tag{5.14}$$

Em outras palavras, a similaridade entre os dois modelos é a maior similaridade observada dentre as Q câmeras.

5.2.3 Atualização de modelos

Mudanças de postura, alterações na iluminação e a dinâmica das oclusões podem modificar a distribuição de cores observada para o objeto ao longo do tempo. Assim, um mecanismo para atualização dos modelos faz-se necessário.

Os modelos são atualizados através de uma combinação linear definida por uma constante de aprendizado α. Se h_{t-1} é um histograma do modelo no instante $t-1$ e h_{obs} é o histograma equivalente (relativo à mesma região e canal), computado a partir da imagem e da máscara M^q, então o modelo é atualizado através de

$$h_t[b] = (1-\alpha)h_{t-1}[b] + \alpha h_{\text{obs}}[b], \tag{5.15}$$

para cada b entre as partições do histograma. O objetivo é preservar a informação do histórico das aparências observado para o objeto ao mesmo tempo em que uma nova informação sobre as distribuições de cor é incorporada.

5.3 Associação

O processo de associação visa atribuir a melhor observação $\mathbf{X}_{t,i}$ possível para cada rastreador $\mathcal{T}_j$. O rastreador $T_j = \langle \hat{\mathrm{s}}_{t-1}, \mathrm{P}_{t-1}, \mathcal{A}, L, C \rangle$ é composto pelas estimações do filtro de Kalman no instante anterior $t-1$, por um modelo de aparência em múltiplas câmeras $\mathcal{A}$, por um *rótulo* L e por um *contador* C.

O rótulo L pode apresentar um dentre quatro valores:

- ATIVO: o valor ATIVO é utilizado para indicar um rastreador que está ativo há algum tempo e, no instante anterior, foi associado a uma observação – logo está rastreando um objeto com sucesso.

- PERDIDO: um rastreador apresenta o rótulo PERDIDO caso ele esteja ativo há algum tempo mas não tenha sido associado à nenhuma observação no instante anterior. Seu estado corrente foi obtido através da predição do filtro de Kalman (a estimação *a posteriori* para a posição é igual à estimação *a priori*).

- NOVO: o rótulo NOVO é utilizado para indicar um rastreador que foi inicializado recentemente – ele precisa ser associado a algumas observações ao longo do tempo antes de ser considerado ativo.

- INATIVO: o rótulo INATIVO é utilizado para indicar um rastreador incapaz de localizar seu objeto por um tempo muito longo – esse rastreador deverá ser removido do processo no próximo instante.

O contador C indica o número de vezes em que o rastreador foi associado a uma observação com sucesso, sendo utilizado para gerenciar a mudança de rótulo.

Dois critérios são avaliados durante a associação:

1. a similaridade entre o modelos de aparências da observação e do objeto rastreado e

2. a posição da observação, avaliada através do filtro do Kalman do rastreador.

A similaridade entre os modelos atua como restrição. Só serão consideradas associações $(\mathbf{X}_{t,i}, \mathcal{T}_j)$ para as quais a similaridade entre os modelos seja igual ou superior a um mínimo estabelecido, isto é, $\text{sim}(\mathcal{A}_{\mathbf{X}_{t,i}}, \mathcal{A}) \geq k_{\text{sim}}$, onde $\mathcal{A}_{\mathbf{X}_{t,i}}$ é o modelo de aparência obtido para $\mathbf{X}_{t,i}$, como descrito na Seção 5.2. Já a avaliação da posição é realizada com o auxílio das estimativas fornecidas pelo filtro de Kalman. Seja $\mathbf{z}_{t,i} = (X, Y, \Delta X, \Delta Y)^\top$ tal que $\mathbf{X}_{t,i} = (X, Y)$ e

$$\begin{pmatrix} \Delta X \\ \Delta Y \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix} - \mathtt{G}\hat{\mathtt{s}}_{t-1}, \tag{5.16}$$

ou seja, $\mathbf{z}_i$ é o estado induzido por $\mathbf{X}_{t,i}$, representando a posição e a velocidade instantânea observadas. Utilizando-se $\mathbf{z}_{t,i}$, é possível definir uma probabilidade para $\mathbf{X}_{t,i}$ condicionada ao rastreador $\mathcal{T}_j$:

$$p(\mathbf{X}_{t,i}|\mathcal{T}_j) = p(\mathbf{z}_{t,\mathtt{i}}|\hat{\mathtt{s}}_t^-, \mathtt{P}_t^-) = \frac{1}{2\pi\|\mathtt{P}_t^-\|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}_{t,\mathtt{i}} - \hat{\mathtt{s}}_t^-)^\top \mathtt{P}_t^{-\,-1}(\mathbf{z}_{t,\mathtt{i}} - \hat{\mathtt{s}}_t^-)\right). \tag{5.17}$$

Essa probabilidade também é utilizada como restrição: só serão consideradas associações $(\mathbf{X}_{t,i}, \mathcal{T}_j)$ para as quais $p(\mathbf{X}_{t,i}|\mathcal{T}_j) \geq k_p$, sendo k_p uma constante estabelecida empiricamente.

A associação é realizada através de uma fila de prioridades. Todas as associações $(\mathbf{X}_{t,i}, \mathcal{T}_j)$ que atendem às restrições descritas anteriormente são inseridas na fila, que determina as prioridades de cada associação de acordo com o seguinte critério:

1. uma associação a um rastreador ATIVO possui prioridade sobre qualquer associação a um rastreador que não esteja ATIVO;

2. uma associação a um rastreador PERDIDO tem prioridade sobre qualquer associação a um rastreador NOVO;

3. se as associações sendo comparadas se referem a rastreadores com o mesmo rótulo, a prioridade é dada à associação que apresenta a maior probabilidade condicional $p(\mathbf{X}_{t,i}|\mathcal{T}_j)$.

A heurística definida pelo critério acima visa priorizar os rastreadores ativos que têm rastreado seu objeto com sucesso. Ela dá precedência aos rastreadores a mais tempo em atividade, evitando que novos rastreadores se estabeleçam, gerando trajetórias fracionadas e incoerências na identificação.

O procedimento é realizado da seguinte forma. A cada iteração, toma-se a associação $(\mathbf{X}_{t,i}, \mathcal{T}_j)$ de maior prioridade. O filtro de Kalman do rastreador $\mathcal{T}_j$ é atualizado de acordo com o procedimento de correção sumarizado na Figura 5.1.1 e seu modelo de aparência é atualizado aplicando-se (5.15) a cada histograma do modelo. Em seguida, todas as associações envolvendo $\mathbf{X}_{t,i}$ ou $\mathcal{T}_j$ são *removidas* da fila, garantindo que a observação não seja associada a nenhum outro rastreador. O procedimento itera até que a fila esteja vazia.

O método gera dois subprodutos que necessitam de atenção: um conjunto de rastreadores para os quais não foi associada nenhuma observação e um conjunto de observações que não foram associadas a um rastreador.

Os rastreadores não associados a observações tem seu estado atualizado a partir do valor da predição por

$$\hat{\mathbf{s}}_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & k_v & 0 \\ 0 & 0 & 0 & k_v \end{bmatrix} \hat{\mathbf{s}}_t^- \tag{5.18}$$

onde $0 < k_v < 1$. O estado é assim projetado para o instante futuro utilizando a informação existente, sem observações, mas reduzindo a velocidade. A diminuição da velocidade é uma heurística que se mostrou útil em experimentos: se o indivíduo rastreado não é observado sucessivamente, não é possível confiar na dinâmica do processo por um longo período, pois pessoas caminhando apresentam padrões erráticos de movimentação. Cada vez que um rastreador não é associado a uma observação seu contador C é decrementado, indicando que ele é menos confiável. C é incrementado sempre que o rastreador é associado a uma observação, até um valor limite. Após um determinado período sem associações no estado PERDIDO, o rastreador se torna não confiável e ele é marcado então como INATIVO.

Para compensar as incertezas decorrentes da perda do objeto (ou não associação com um $\mathbf{X}_{t,i}$), a covariância do erro é ampliada de acordo com

$$\mathrm{P}_t = \begin{bmatrix} k_l & 0 & 0 & 0 \\ 0 & k_l & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathrm{P}_t^- \tag{5.19}$$

sendo $k_l > 1$. Intuitivamente, a área de busca determinada pelo filtro será ampliada, na tentativa de localizar o objeto no próximo instante.

As observações não associadas a nenhum rastreador dão origem a novos rastreadores. Esses rastreadores recebem o rótulo NOVO e têm seus filtros de Kalman inicializados com estados na posição da observação e velocidade igual a zero. Após um certo período, caso esse rastreador NOVO se mostre confiável, ele é promovido a ATIVO.

Capítulo 6

Resultados

Neste capítulo serão apresentados alguns resultados provenientes de implementações e testes dos métodos propostos. Para os testes, foram escolhidas as bases de dados do PETS 2006 [48] e do PETS 2009 [18]. Essas bases foram escolhidas pois são públicas e representam situações desafiadoras em cenas realistas.

A Seção 6.1 exibe os resultados obtidos para a subtração de fundo a partir do método de mistura de Gaussianas apresentado no Capítulo 3. A estimação de matrizes de homografia é brevemente comentada na Seção 6.2. Resultados de detecção instantânea de indivíduos são exibidos na Seção 6.3. Finalmente, resultados para o rastreamento de múltiplos indivíduos em uma sequencia do PETS 2006 é apresentada na Seção 6.4.

6.1 Subtração de fundo

A solução adotada para a subtração de fundo foi modelar a distribuição de cor de cada pixel por uma mistura de Gaussianas, como descrito no Capítulo 3. O software Cluster, implementado por Bouman [6], foi realizado para o treinamento dos modelos. Cluster combina o método de maximização de esperança (EM) com um algoritmo k-médias [15] – este último empregado para inferir o número de Gaussianas utilizado pelo modelo de mistura.

A Figura 6.1.1 exibe resultados de classificação de fundo para 3 câmeras da sequencia S07 do PETS 2006 [48]. A própria sequencia foi utilizada como treinamento para o algoritmo EM. Até 3 Gaussianas foram utilizadas para modelar o fundo, escolhidas dentre as distribuições

de maior peso encontradas pelo algoritmo EM. Gaussianas que apresentassem peso inferior a 20% foram descartadas. Os mesmos parâmetros foram utilizados para as três câmeras: a fronteira de classificação $\alpha = 1.5$ e a constante $\beta = 0.7$, utilizada pelo método de remoção de sombras.

A Figura 6.1.2 exibe resultados de classificação de fundo para 6 câmeras da sequencia S2L1 do PETS 2009 [18]. Um conjunto de treinamento que constitui a base de dados foi utilizado com o algoritmo EM. Novamente, até 3 Gaussianas foram utilizadas para modelar o fundo e Gaussianas que apresentassem peso inferior a 20% foram descartadas. A fronteira de classificação $\alpha = 5.0$ e a constante $\beta = 0.7$ foram utilizadas em todas as câmeras, exceto pela câmera 2 na qual, para reduzir o número de falsos-positivos, teve o valor α incrementado para 7.0. Essa base de dados é mais desafiadora em relação à subtração de fundo por ser composta de cenas externas, onde a variação de luz e sombra é muito maior. As diferenças de iluminação existentes entre a sequencia testada e a sequencia de treinamento fazem com que os prédios vistos nas câmeras 1 e 2 apresentem regiões classificadas erroneamente como figura. Métodos adaptativos de subtração de fundo, como o proposto por Stauffer e Grimson [47] podem ser uma opção mais adequada nessas situações.

6.1.1 Erros de subtração de fundo

Embora o métodos de detecção de objetos (pessoas) que será apresentado a seguir mostre-se robusto a erros de subtração fundo, tais falhas de classificação podem prejudicar a detecção, principalmente se tratando de falsos-negativos. A Figura 6.1.3 ilustra os erros mais comuns: sombras apresentando alto contraste e camuflagem (discutida no Capítulo 3).

6.2 Homografias

As homografias utilizadas foram obtidas pelo método de estimação homogênea [24], apresentado no Capítulo 3. Foi utilizada a implementação Octave do método desenvolvida disponibilizada por Kovesi [35]. Os pontos de correspondência foram obtidos através de calibração

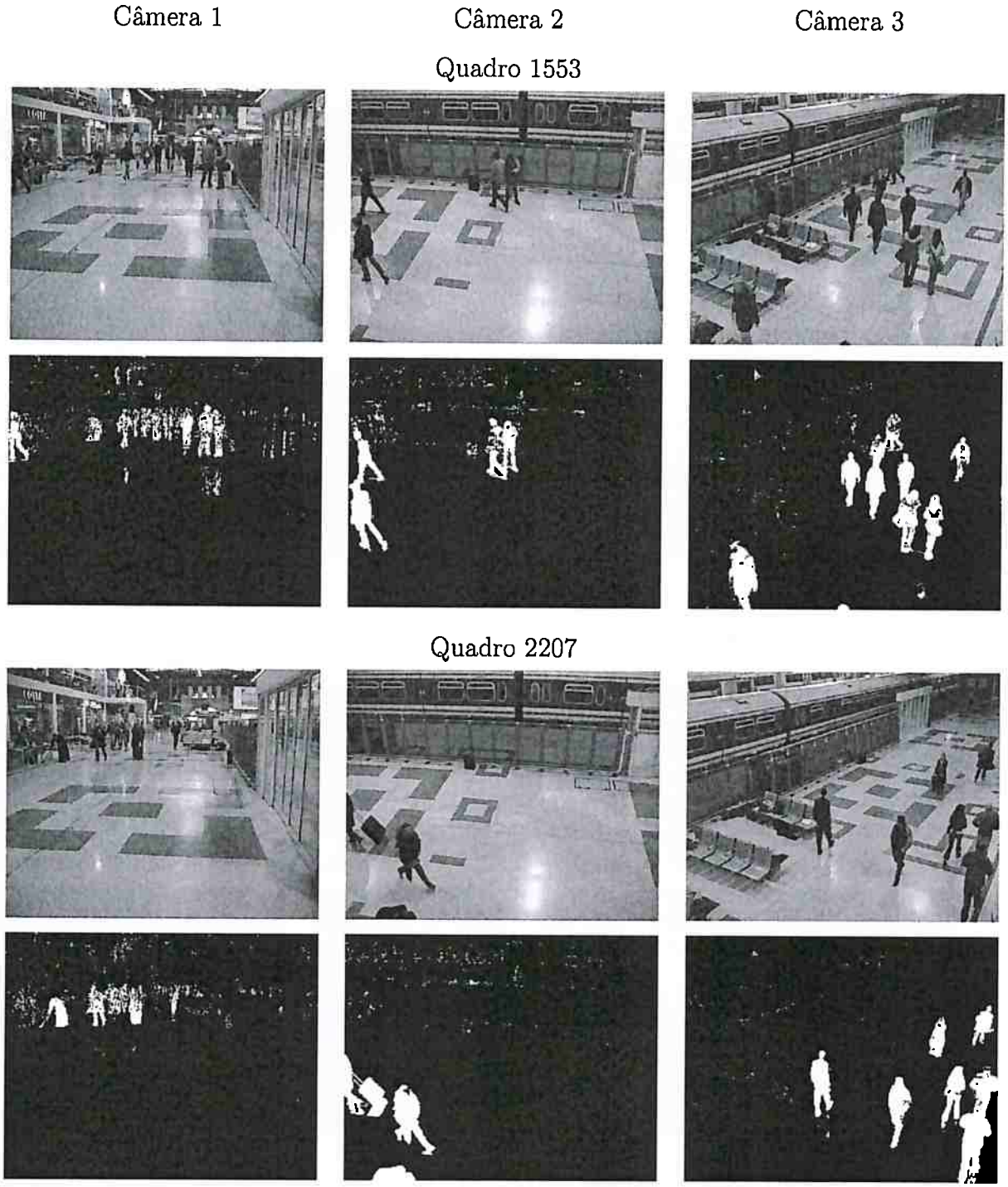


Figura 6.1.1: **Subtração de fundo – PETS 2006.** Subtração de fundo em dois instantes da sequencia S07 do PETS 2006 [48]. Foram utilizadas até 3 Gaussianas no modelo (um peso mínimo de 20% foi exigido para que a Gaussiana fosse aceita como modelo de fundo. A fronteira de decisão foi determinada por $\alpha = 1.5$. Sombras foram identificadas utilizando-se $\beta = 0.7$.

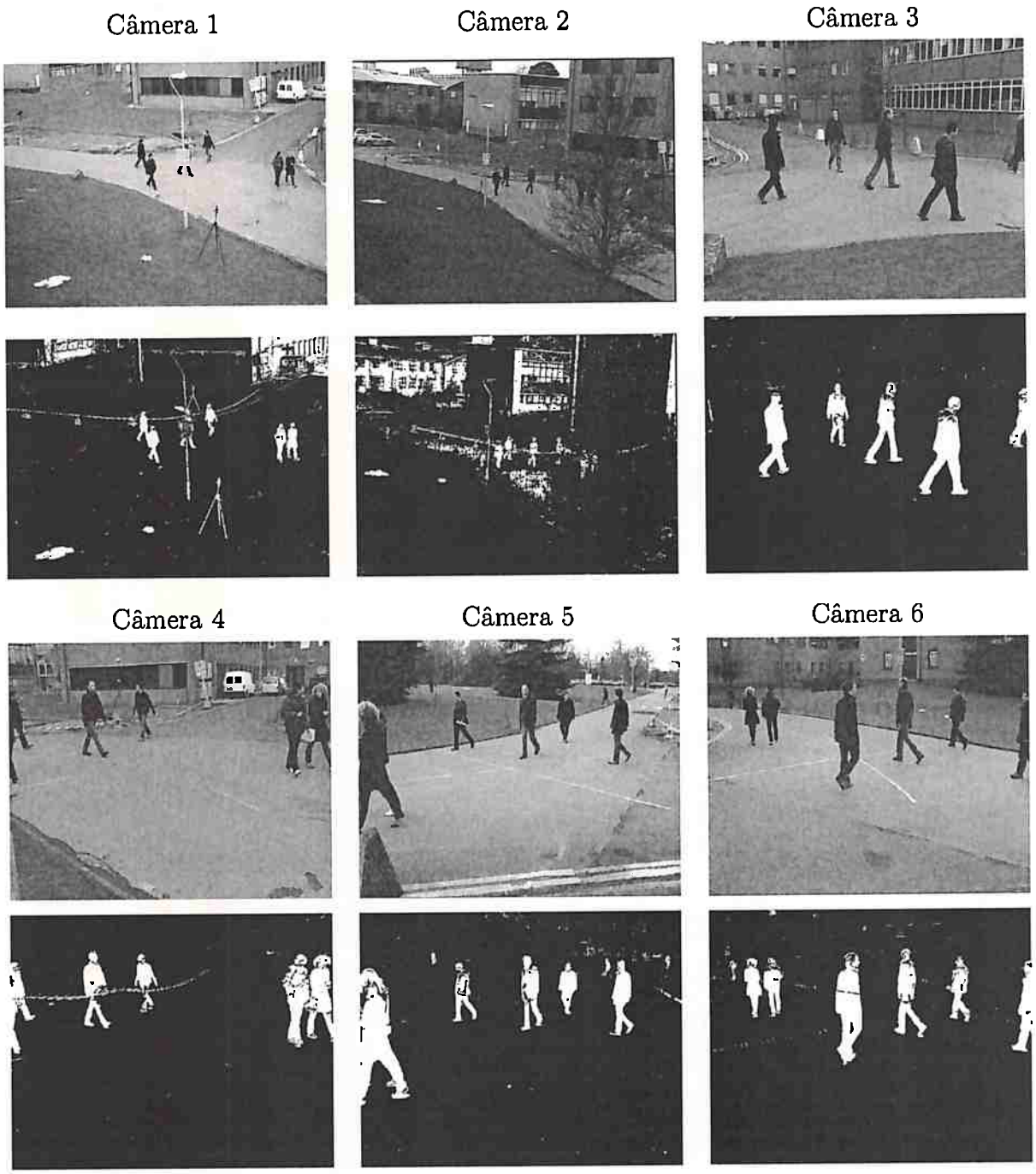


Figura 6.1.2: **Subtração de fundo – PETS 2009**. Subtração de fundo para o quadro 340 da sequencia S2L1 do PETS 2009 [18]. Foram utilizadas até 3 Gaussianas no modelo (um peso mínimo de 20% foi exigido para que a Gaussiana fosse aceita como modelo de fundo. A fronteira de decisão foi determinada por $\alpha = 5.0$ (exceto pela câmera 2, onde foi utilizado $\alpha = 7.0$). Sombras foram identificadas utilizando-se $\beta = 0.7$. Diferenças de iluminação com a sequencia de treinamento fazem com que as paredes dos prédios sejam classificadas como figura nas câmeras 1 e 2.

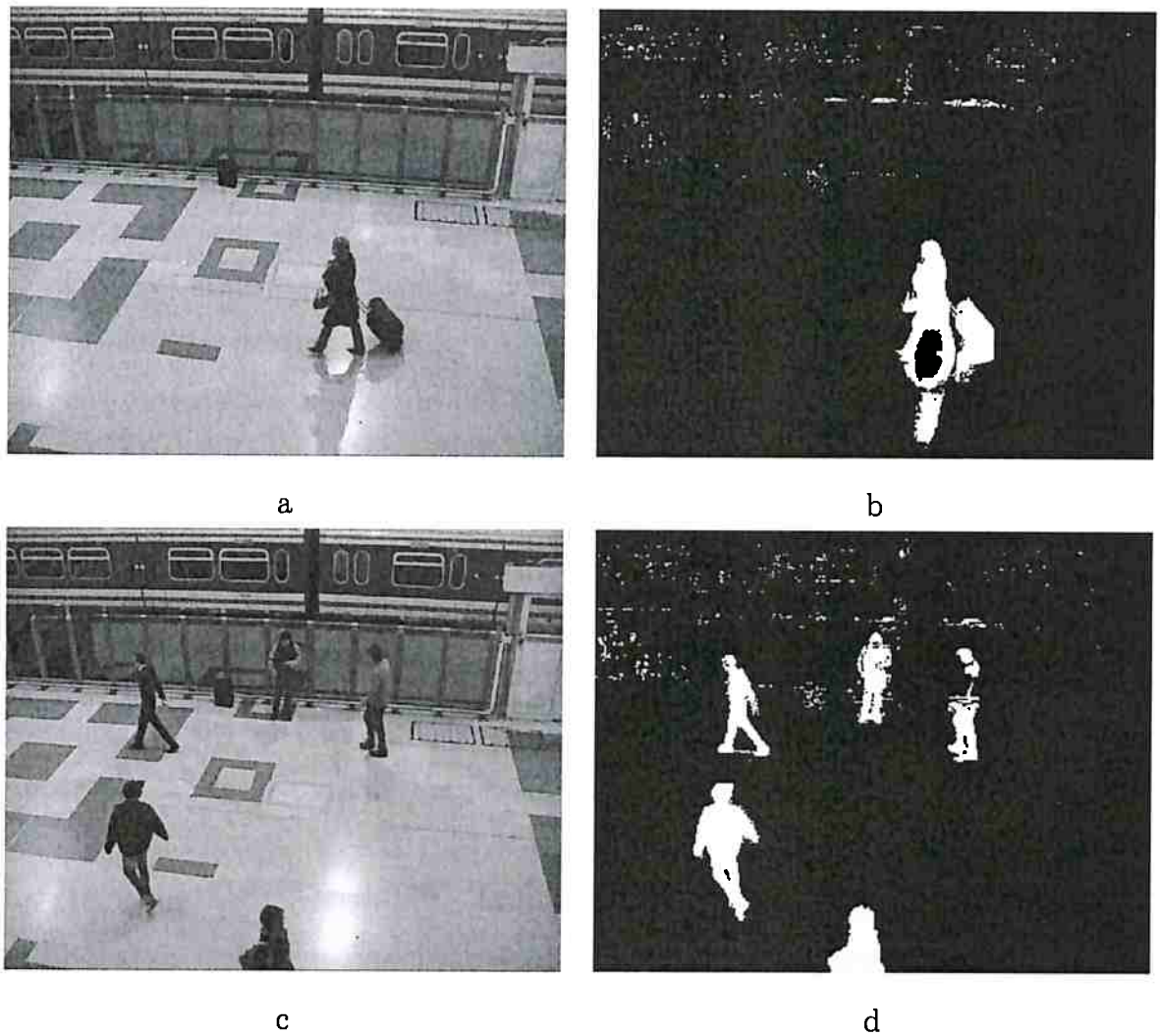


Figura 6.1.3: **Problemas comuns em subtração de fundo.** Dois exemplos de problemas comumente encontrados em subtração de fundo. (a) Sombra apresentando alto contraste produz falsos-positivos (b). (c) Homem trajando roupas com distribuição de cor similar à do fundo (camuflagem) produz falsos-negativos (d). Camuflagem é o erro de classificação que mais interfere no desempenho do método de detecção de objetos.

disponibilizada pelas bases de dados utilizadas (PETS).

6.3 Detecção

As Figuras 6.3.1, 6.3.2 e 6.3.3 exibem resultados para detecção obtidos para a base de dados do PETS 2006 (sequencia S07) enquanto que as Figuras 6.3.4 e 6.3.5 se referem à base do PETS 2009 (sequencia S2L1). Os pontos vermelhos indicam as localizações obtidas pelo sistema, exibidas tanto no plano de solo quanto nos planos de imagem para referência. A detecção é instantânea, utilizando apenas o Algoritmo 1 proposto no Capítulo 4, sem fazem uso de qualquer outro método para obter consistência temporal (rastreamento, por exemplo).

6.4 Rastreamento

Anotação de referência (*ground-truth*) foi produzida manualmente para avaliar o desempenho do algoritmo de rastreamento. A posição de cada indivíduo foi anotada para 150 quadros, amostrados em um intervalo regular de 10 quadros dentre os 1500 existentes na sequencia S07 do PETS 2006. Os indivíduos forma identificados de forma consistente para que as trajetórias de cada indivíduo pudessem ser avaliadas.

A Tabela 6.1 apresenta os resultados obtidos. Todos os 22 indivíduos existentes na sequencia forma associados com sucesso a uma ou mais trilhas produzidas pelo sistema. Só uma única trilha gerada pelo módulo de rastreamento não pôde ser associada a nenhum indivíduo.

Idealmente, um único rastreador deveria acompanhar o mesmo indivíduo ao longo de todo o vídeo. O sistema proposto produziu uma média de 1.32 trajetórias por indivíduo, o que corresponde a poucos erros durante o rastreamento. Ao longo de toda a sequencia, só houve um único caso de troca de trajetória, ou seja, um caso em que o sistema cometeu um erro de associação, invertendo a trajetória de dois indivíduos. Esta troca ocorreu entre dois indivíduos vistos por apenas duas câmeras, em oclusão e alinhados a linha base entre os centros de projeção das câmeras (*baseline*).

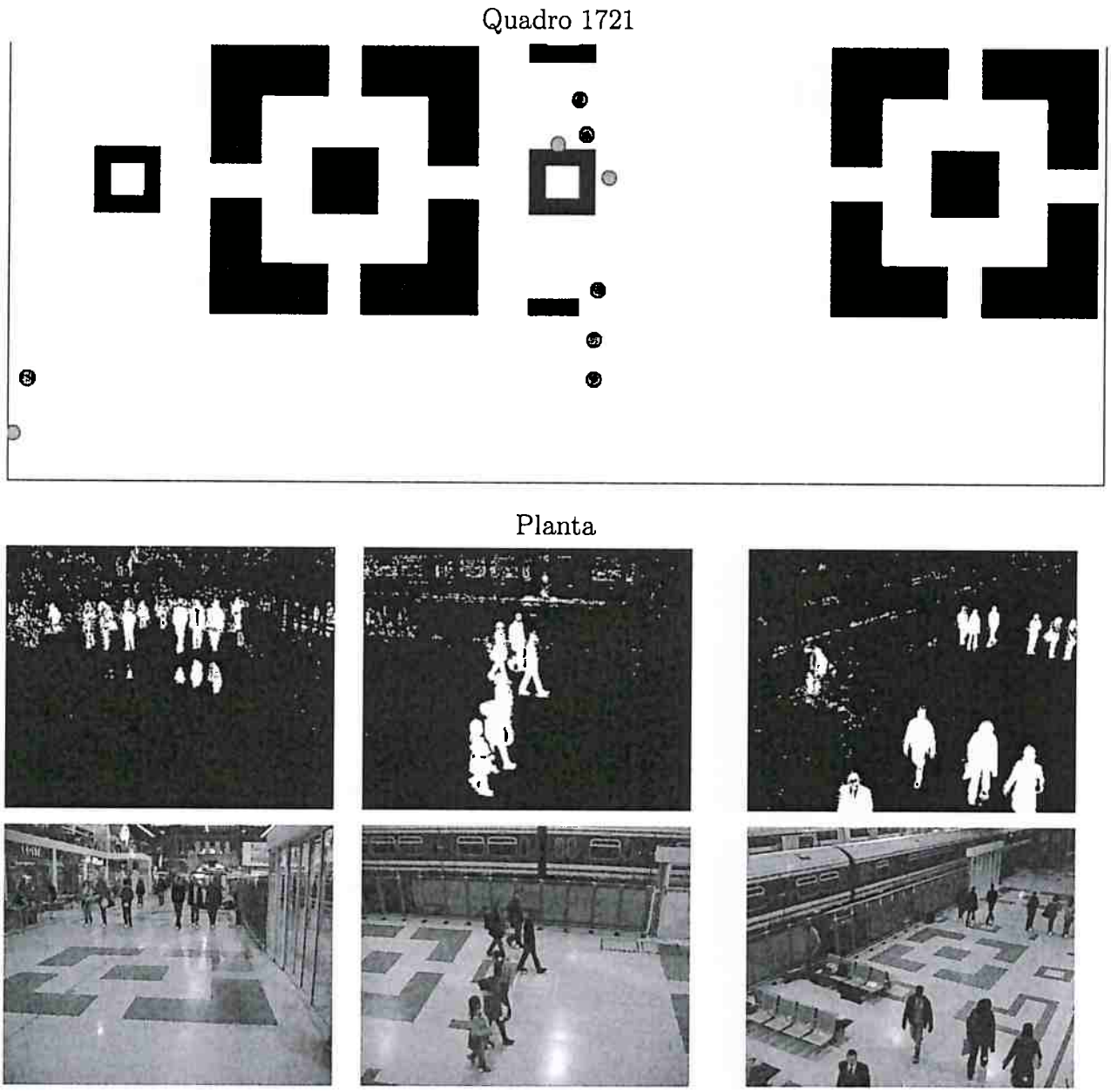


Figura 6.3.1: **Resultados PETS 2006, quadro 1721.**

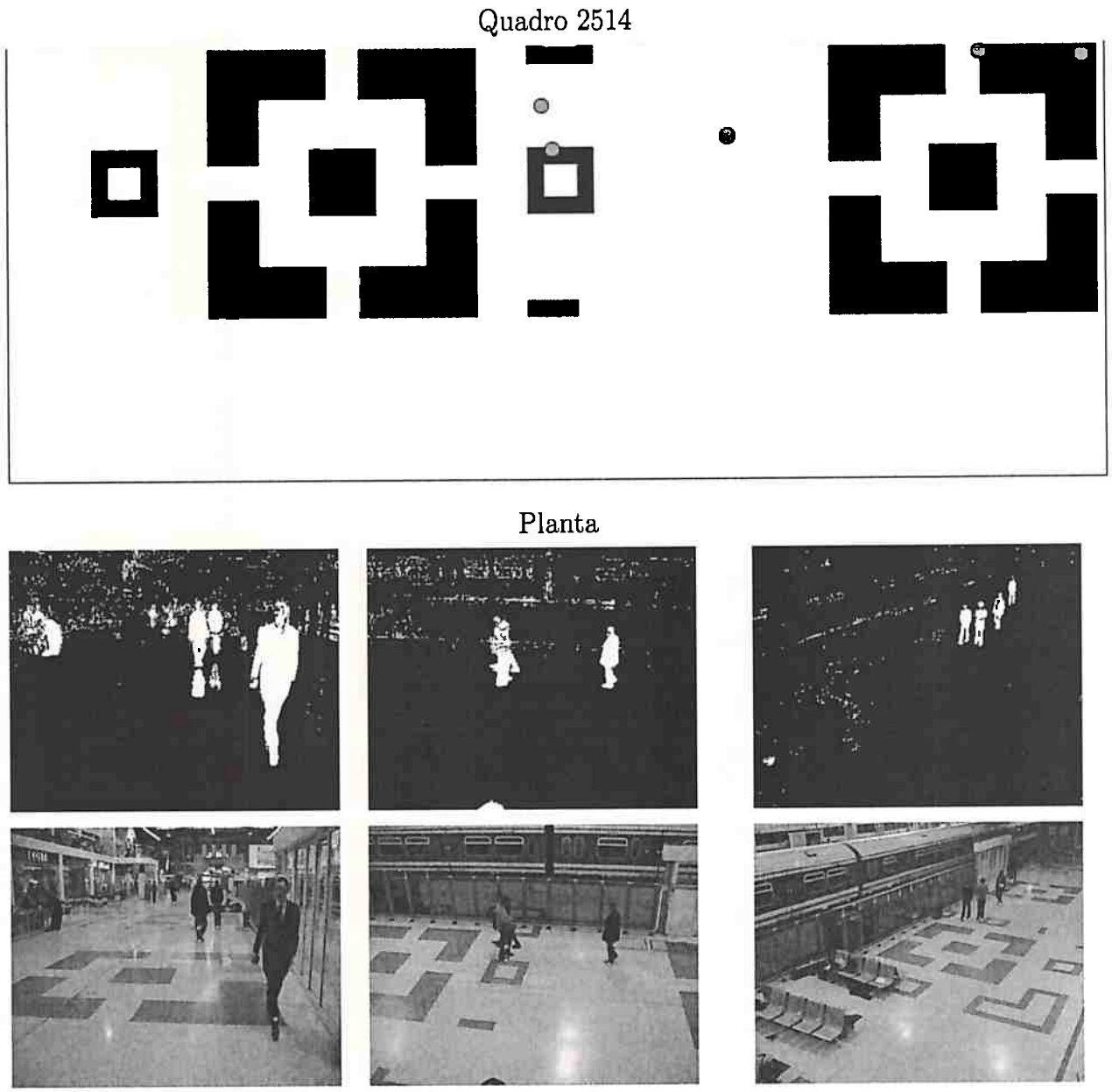


Figura 6.3.2: **Resultados PETS 2006, quadro 2514.**

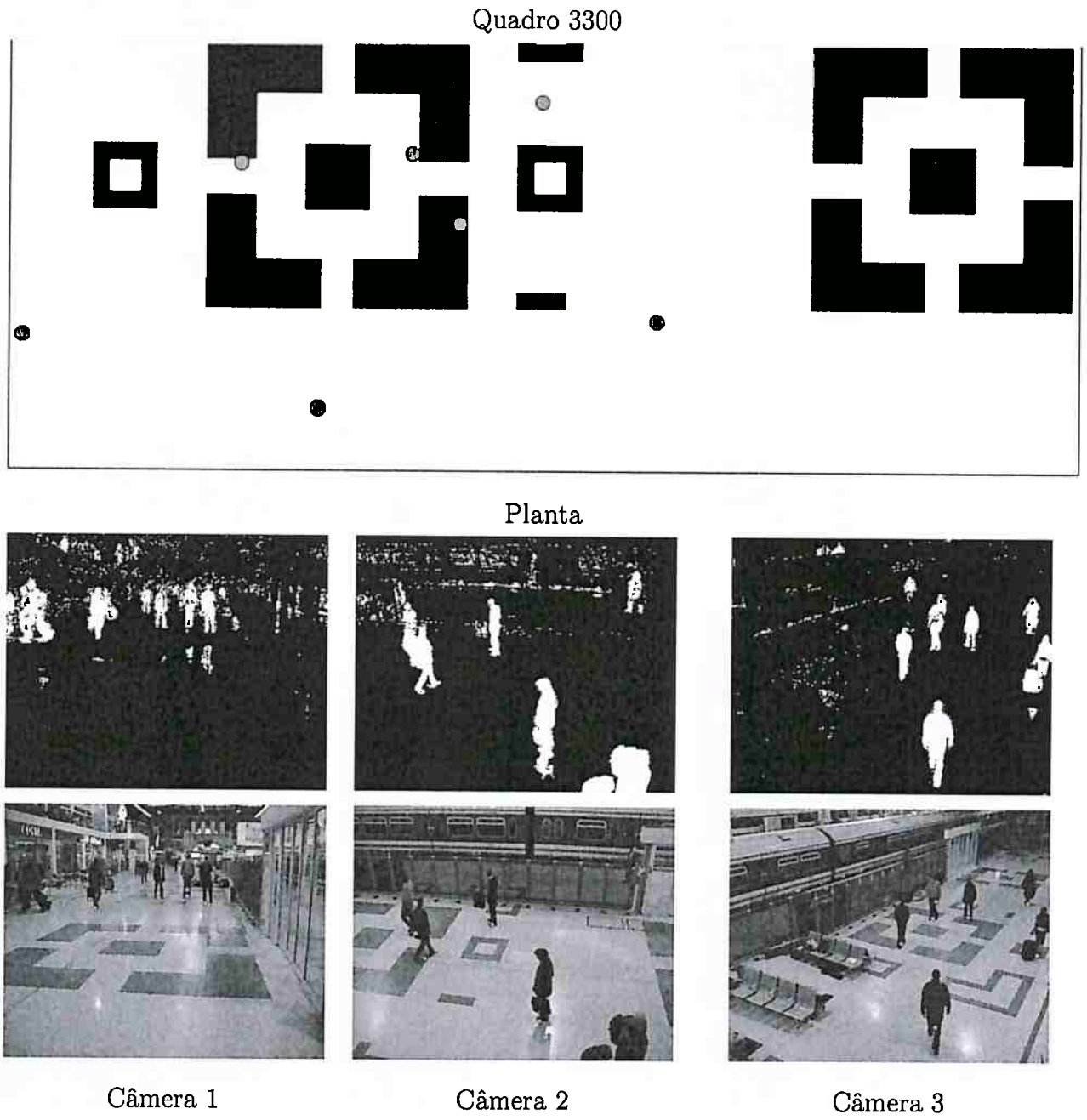


Figura 6.3.3: **Resultados PETS 2006, quadro 3300.**

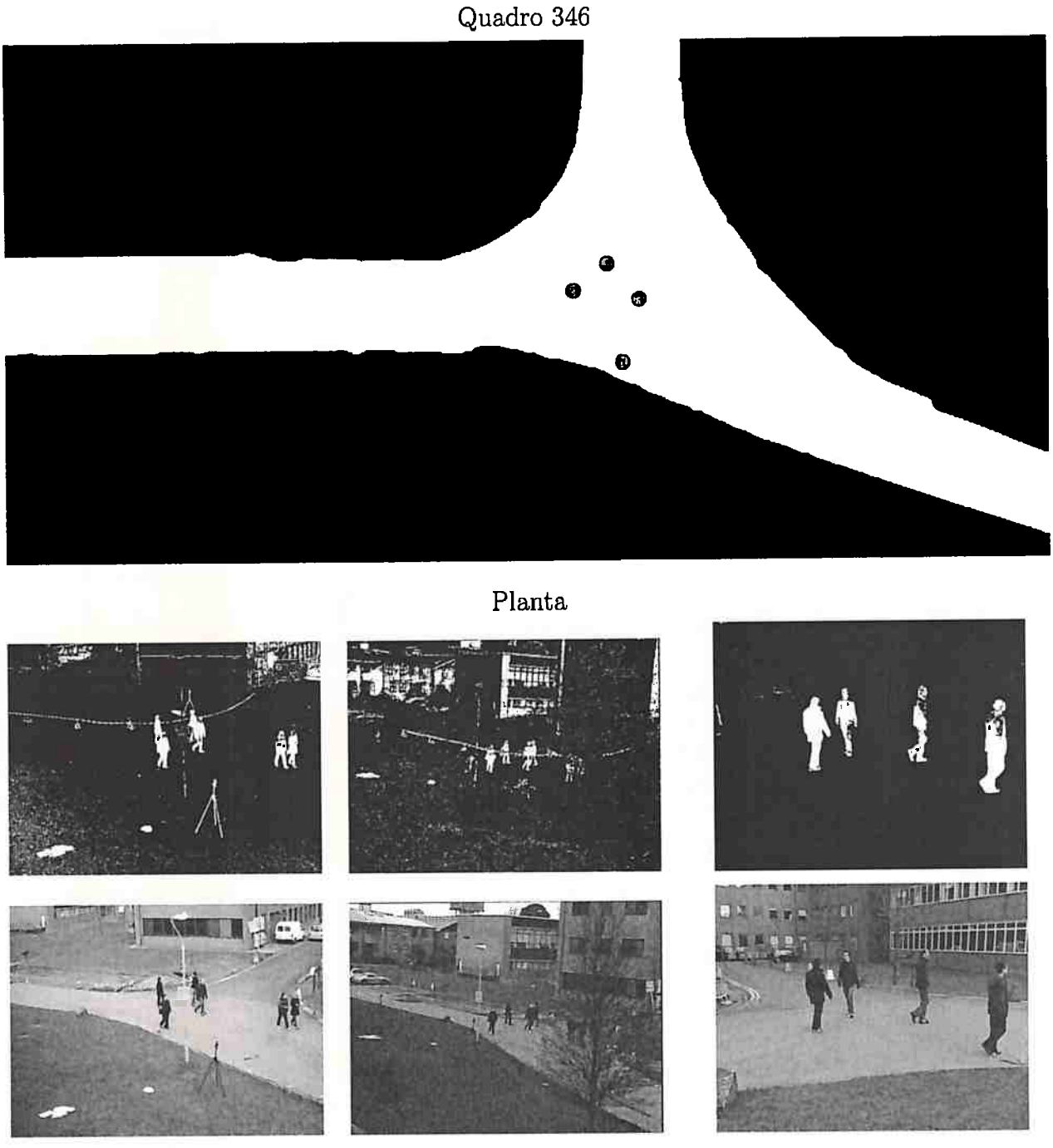


Figura 6.3.4: **Resultados PETS 2009, quadro 346 S2L1.**

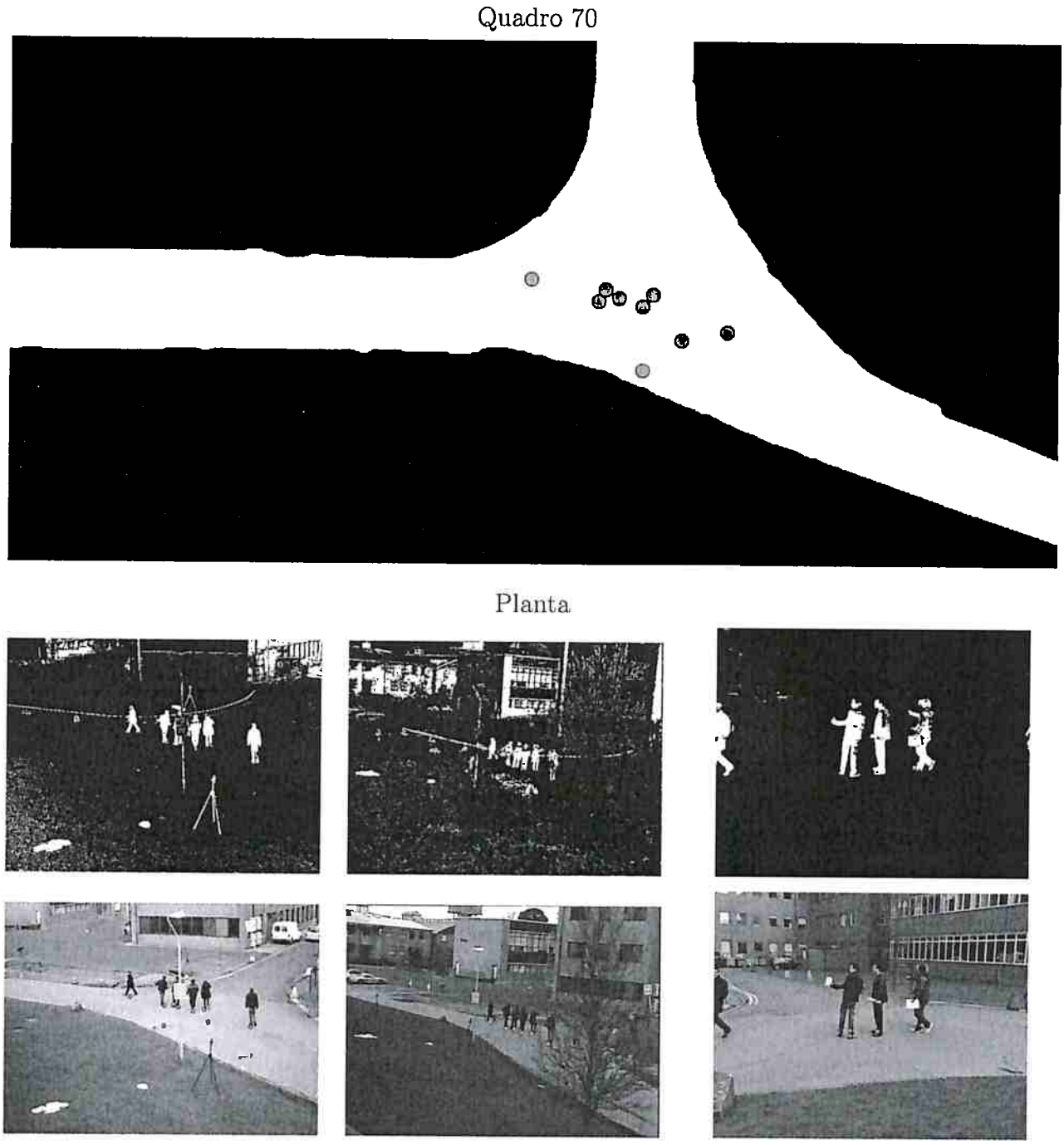


Figura 6.3.5: **Resultados PETS 2009, quadro 70 S2L1.**

	PETS 2006 S07
Número de indivíduos	22
Trajetórias encontradas	30
Cobertura da trajetória	100.00%
Precisão	96.67%
Trajetórias por indivíduo	1.3182

Tabela 6.1: **Resultados obtidos para rastreamento em PETS 06 S07**. Comparação entre os resultados obtidos pelo rastreador e a anotação de referência.

A Figura 6.4.1 exibe a raiz do desvio quadrado médio entre as posições estimadas pelas trajetórias e as posições observadas na anotação de referência para cada indivíduo. O maior desvio observado foi de cerca de 50 cm, associado a um indivíduo que aparece correndo na cena (indivíduo 14). Já a Figura 6.4.2 exibe a trajetória encontrada pelo sistema e a trajetória de referência para o indivíduo 19. Esse indivíduo cruza todo o saguão da estação e é ocluso por outras pessoas em diversas ocasiões.

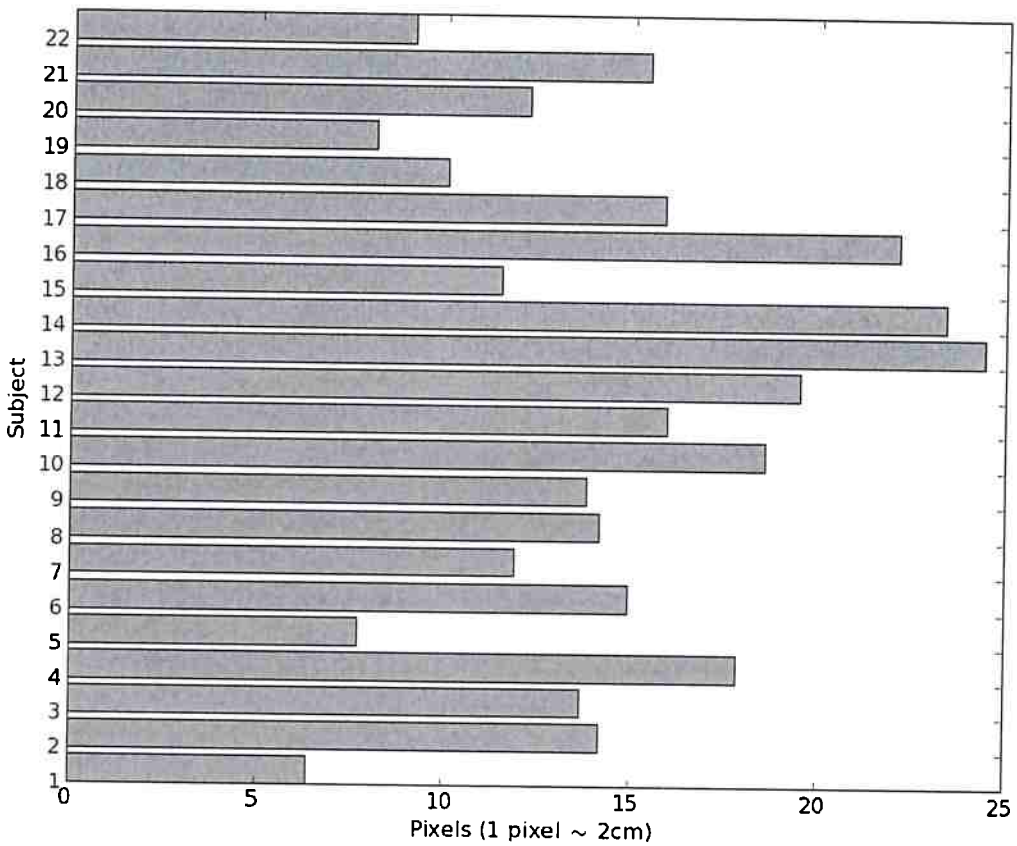


Figura 6.4.1: **Raiz do desvio quadrado médio**. A raiz do desvio quadrado médio observado no rastreamento de cada indivíduo é exibida para a sequencia S07 do PETS 2006. Cada pixel no plano do solo equivale a cerca de 2cm. O maior erro de localização não excedeu 50cm, causado por um indivíduo correndo pelo saguão.

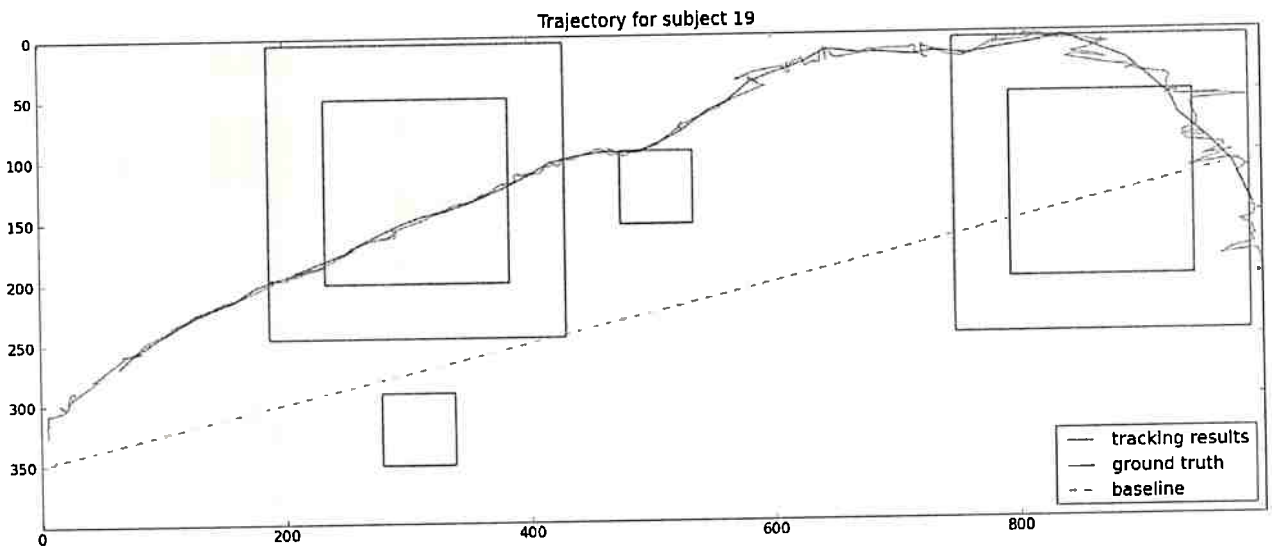


Figura 6.4.2: **Trajetória observada para o indivíduo 19.** O indivíduo esteve ocluso diversas vezes ao longo de sua trajetória pelo saguão. Há falsos-negativos em sua subtração de fundo devido à camuflagem (a camisa do indivíduo tem cores similares ao fundo da cena).

Capítulo 7

Conclusões

Para se tornarem largamente empregados em aplicações diversas, sistemas de visão computacional precisam se tornar robustos frente às situações cotidianas que os usuários vivem em seus ambientes de trabalho.

Quando utilizados para armazenamento de dados e organização de informação, sistemas computacionais costumam impressionar positivamente seus usuários, que frequentemente têm a impressão que a capacidade das máquinas nessas tarefas excede suas própria capacidade humana de memorização. O inverso ocorre em relação à visão computacional: dada a grande eficiência com que o sistema visual humano lida com tarefas extremamente complexas de sensoreamento, a experiência dos usuários com sistemas de visão computacional é geralmente decepcionante. "Por que o computador não consegue me ver?", é um questionamento comum, já que a habilidade é dada como certa e natural entre seres humanos.

O problema abordado nesta tese é um exemplo de tal situação. Dentro de certos limites de complexidade, a visão humana é extremamente hábil em detectar e rastrear indivíduos, mesmo sob oclusão. Tal habilidade é tão natural às pessoas que torna-se imediata a necessidade que sistemas computacionais sejam capazes de obter níveis próximos de eficácia, com a intenção de prover uma gama de serviços e aplicações.

Na tentativa de obter robustez no rastreamento de múltiplos indivíduos em ambientes diversos, o uso de múltiplos sensores tornou-se uma alternativa comum. Como visto no Capítulo 2, várias técnicas foram desenvolvidas utilizando redes de câmeras sincronizadas.

Como resposta ao problema inerente de integração entre esses múltiplos sensores, a restrição homográfica, induzida pelo plano no qual os objetos trafegam, foi empregada com sucesso por diversos autores. Contudo, vários autores [33, 27] buscam segmentar as componentes conexas das figuras em indivíduos. Isto significa que um problema tão ou mais difícil precisa ser abordado anteriormente: segmentação de imagens.

Os métodos vistos na presente tese se assemelham aos trabalhos de Khan e Shah ?? e Eshel e Moses [17], que tratam os problemas de segmentação e detecção de forma conjunta, integrando a informação de várias câmeras antes de impor qualquer hipótese à disposição dos indivíduos em cena.

A contribuição deste trabalho está em postergar o problema de segmentação e tratar a detecção como um processo de acúmulo de evidência. As evidências são os pixeis de figura observados pelas várias câmeras da rede de sensores. Combinadas a propriedades geométricas como pontos ideais e a razão cruzada, essas evidências podem ser combinadas na formação de hipóteses verossímeis, capazes de lidar com oclusões. A integração de evidência coletada por múltiplas câmeras é realizada através da restrição homográfica. Todo o processo pode ser tratado dentro do bem conhecido arcabouço da Transformada de Hough, utilizada há anos como mecanismo de integração de evidência e detecção de estruturas, como discutido no Capítulo 4. Este arcabouço é capaz de tratar redes de câmeras de tamanhos arbitrários: as evidências coletadas por cada nova câmera são facilmente integradas ao sistema, contribuindo com a detecção dos objetos. Diversos resultados obtidos a partir de bases de dados públicas demonstram a viabilidade e eficácia do método.

O problema é transformado para um espaço que pode ser entendido como uma planta do local sensoreado. Nesse espaço, uma "vista aérea" da cena, os objetos em trânsito podem ser vistos de forma esquemática e livre de oclusão, o que é particularmente adequado ao rastreamento e à análise de trajetórias. Auxiliado por um modelo de aparência capaz de modelar a distribuição de cor nas regiões do torso e das pernas de cada indivíduo, construído a partir das imagens provenientes de cada câmera, o sistema proposto no Capítulo 5 é capaz de atribuir observações obtidas na fase de detecção a Filtros de Kalman que realizam os

rastreamento dos indivíduos. Este método para o gerenciar o rastreamento de múltiplos objetos é outra contribuição da presente tese.

O sistema apresentado foi testado em situações desafiadoras em bases de dados públicas e apresentou bom desempenho na localização e rastreamento de diversos indivíduos, mesmo sob oclusão mútua.

7.1 Trabalho futuro

O trabalho pode ser ampliado de várias formas. Alguns tópicos para trabalho futuro são:

- Maior integração entre as componentes de subtração de fundo, detecção e rastreamento. Os modelos de aparência podem ser aplicados à subtração de fundo como forma de tratar casos de falsos-negativos – um modelo de aparência, aliado ao rastreamento, pode ser empregado no tratamento de casos de camuflagem.

- Estimação automática de homografias. A integração dos vários sensores é feita por homografias que são obtidas através de correspondências entre pontos, que são fornecidas manualmente quando não há calibração total das câmeras (caso no qual as homografias podem ser algebricamente derivadas).

- Análise do erro encontrado na estimação da localização de cada indivíduo e como esse erro pode variar para diferentes câmeras em diferentes regiões do plano de imagem.

- Extensões para situações nas quais a superfície em que os objetos de interesse trafegam não é plana.

Finalmente, sistemas para detecção de *eventos* em cenas podem ser desenvolvidos sobre o arcabouço de detecção e rastreamento proposto aqui. Trajetórias, relações de oclusão e as alturas estimadas para os indivíduos podem ser empregadas na caracterização de eventos e situações de interesse em várias aplicações.

Referências Bibliográficas

[1] Ballard, D. H. *Generalizing the hough transform to detect arbitrary shapes.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, pp. 714–725. Citado na(s) página(s) 43

[2] Bar-Shalom, Y., Li, X. R., and Kirubarajan, T. *Estimation with Applications to Tracking and Navigation*, 1 ed. Wiley-Interscience, June 2001. Citado na(s) página(s) 82

[3] Barnard, S. Interpreting Perspective Images. *Artificial Intelligence 21* (1983), 435–462. Citado na(s) página(s) 35

[4] Blackman, S. S. Multiple hypothesis tracking for multiple target tracking. *Aerospace and Electronic Systems Magazine, IEEE 19*, 1 (2004), 5–18. Citado na(s) página(s) 81, 82

[5] Boult, T., Gao, X., Micheals, R., and Eckmann, M. Omni-directional visual surveillance. *Image and Vision Computing 22*, 7 (July 2004), 515–534. Citado na(s) página(s) 24, 25

[6] Bouman, C. A. Cluster: An unsupervised algorithm for modeling Gaussian mixtures. Available from http://www.ece.purdue.edu/~bouman, April 1997. Citado na(s) página(s) 25, 27, 28, 93

[7] BRESENHAM, J. E. Algorithm for computer control of a digital plotter. 1–6. Citado na(s) página(s) 55

[8] CHENG, Y. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 17*, 8 (1995), 790–799. Citado na(s) página(s) 70, 71

[9] COMANICIU, D., AND MEER, P. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*, 5 (2002), 603–619. Citado na(s) página(s) 70

[10] COMANICIU, D., RAMESH, V., AND MEER, P. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 25*, 5 (2003), 564–577. Citado na(s) página(s) 70

[11] CRIMINISI, A., REID, I. D., AND ZISSERMAN, A. Single view metrology. *International Journal of Computer Vision 40*, 2 (2000), 123–148. Citado na(s) página(s) 31, 51, 57

[12] CROWLEY, J. L. Social perception. *Queue 4*, 6 (2006), 34–43. Citado na(s) página(s) 4, 10

[13] DOUCET, A., DE FREITAS, N., AND GORDON, N., Eds. *Sequential Monte Carlo Methods in Practice (Statistics for Engineering and Information Science)*, 1 ed. Springer, June 2001. Citado na(s) página(s) 11

[14] DUDA, R. O., AND HART, P. E. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM 15*, 1 (January 1972), 11–15. Citado na(s) página(s) 42, 60

[15] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000. Citado na(s) página(s) 14, 25, 93

[16] ELFES, A. *Occupancy Grids: A probabilistic framework for robot perception and navigation.* PhD thesis, Carnegie-Mellon University, Maio 1989. Citado na(s) página(s) 46

[17] ESHEL, R., AND MOSES, Y. Homography based multiple camera detection and tracking of people in a dense crowd. In *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)* (Los Alamitos, CA, USA, 2008), pp. 1–8. Citado na(s) página(s) 20, 38, 108

[18] FERRYMAN, J., AND SHAHROKNI, A. An Overview of the PETS 2009 Challenge. In *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2009* (Junho 2009), J. M. Ferryman, Ed., IEEE, pp. 25–30. Citado na(s) página(s) 9, 26, 27, 28, 93, 94, 96

[19] FLEURET, F., BERCLAZ, J., LENGAGNE, R., AND FUA, P. Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 30*, 2 (2008), 267–282. Citado na(s) página(s) 16, 46

[20] FUKUNAGA, K., AND HOSTETLER, L. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on 21*, 1 (1975), 32 40. Citado na(s) página(s) 70

[21] G., S. J., AND KNEEBONE, G. T. *Algebraic Projective Geometry.* Oxford University Press, Novembro 1998. Citado na(s) página(s) 33

[22] GOLUB, G. H., AND VAN LOAN, C. F. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences).* The Johns Hopkins University Press, October 1996. Citado na(s) página(s) 41

[23] HARITAOGLU, I., HARWOOD, D., AND DAVIS, L. S. W^4: real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 8 (Agosto 2000), 809–830. Citado na(s) página(s) 12, 13, 15, 20, 46, 84

[24] HARTLEY, R., AND ZISSERMAN, A. *Multiple View Geometry in Computer Vision.* Cambridge University Press, March 2004. Citado na(s) página(s) 17, 30, 31, 32, 33, 35, 40, 41, 58, 94

[25] HARTLEY, R. I. Theory and practice of projective rectification. *International Journal of Computer Vision 35*, 2 (November 1999), 115–127. Citado na(s) página(s) 71, 73

[26] HOUGH, P. Method and means for recognizing complex patterns. U.S. Patent 3.069.654, Dezembro 1962. Citado na(s) página(s) 41, 42

[27] HU, W., HU, M., ZHOU, X., TAN, T., LOU, J., AND MAYBANK, S. Principal axis-based correspondence between multiple cameras for people tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 4 (2006), 663–671. Citado na(s) página(s) 19, 20, 38, 108

[28] ILLINGWORTH, J., AND KITTLER, J. A survey of the hough transform. *Computer Vision, Graphics and Image Processing 44*, 1 (1988), 87–116. Citado na(s) página(s) 41, 45, 46, 60

[29] ISARD, M., AND BLAKE, A. Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision 29*, 1 (August 1998), 5–28. Citado na(s) página(s) 11

[30] ISARD, M., AND MACCORMICK, J. BraMBLe: a Bayesian multiple-blob tracker. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2001), vol. 2, pp. 34–41 vol.2. Citado na(s) página(s) 14

[31] KHAN, S., AND SHAH, M. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Lecture Notes in Computer Science.* Springer Berlin / Heidelberg, 2006, pp. 133–146. Citado na(s) página(s) 17, 38

[32] KHAN, S., AND SHAH, M. Tracking multiple occluding people by localizing on multiple scene planes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*, 3 (2009), 505–519. Citado na(s) página(s) 17, 38

[33] KIM, K., AND DAVIS, L. Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In *Proceedings of 9th European Conference on Computer Vision (ECCV'06)* (Graz, Áustria, 2006), vol. 3953, pp. 98–109. Citado na(s) página(s) 19, 20, 38, 108

[34] KITAGAWA, G. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics 5*, 1 (1996), 1–25. Citado na(s) página(s) 11

[35] KOVESI, P. D. MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia. Disponível em: <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>. Citado na(s) página(s) 94

[36] MAYBECK, P. S. *Stochastic models, estimation, and control*, vol. 141 of *Mathematics in Science and Engineering*. Academic Press, 1979. Citado na(s) página(s) 11, 82

[37] MCLEAN, G. F., AND KOTTURI, D. Vanishing point detection by line clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 17*, 11 (1995), 1090–1095. Citado na(s) página(s) 35

[38] MITTAL, A., AND DAVIS, L. S. M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. *International Journal of Computer Vision 51*, 3 (February 2003), 189–203. Citado na(s) página(s) 15, 20, 38, 84

[39] OVER, P., AWAD, G., ROSE, T., AND FISCUS, J. TRECVID 2008 Goals, Tasks, Data, Evaluation Mechanisms and Metrics. Tech. rep., National Institute of Standards and Technology – NIST, Abril 2009. Citado na(s) página(s) 9

[40] ROSENFELD, A. Picture processing by computer. *ACM Comput. Surv. 1*, 3 (1969), 147–176. Citado na(s) página(s) 42

[41] SANTOS, T. T., AND MORIMOTO, C. H. People detection under occlusion in multiple camera views. In *Proceedings of XXI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI '08)* (Los Alamitos, Oct. 12–15, 2008 2008), IEEE Computer Society, pp. 53–60. Citado na(s) página(s) 38, 53

[42] SANTOS, T. T., AND MORIMOTO, C. H. Multiple camera people detection and tracking using support integration. *Pattern Recognition Letters* (submetido em 2009). (em revisão). Citado na(s) página(s) 53

[43] SCHAPIRE, R. E., AND SINGER, Y. Improved boosting algorithms using confidence-rated predictions. *Machine Learning 37*, 3 (Dezembro 1999), 297–336. Citado na(s) página(s) 15

[44] SENIOR, A., HAMPAPUR, A., TIAN, Y.-L., BROWNA, L., PANKANTIA, S., AND BOLLE, R. Appearance models for occlusion handling. *Image and Vision Computing 24*, 11 (November 2006), 1233–1243. Citado na(s) página(s) 14, 84

[45] SHI, J., AND MALIK, J. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 8 (2000), 888–905. Citado na(s) página(s) 19

[46] SHUFELT, J. A. Performance evaluation and analysis of vanishing point detection techniques. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 21*, 3 (1999), 282–288. Citado na(s) página(s) 35

[47] STAUFFER, C., AND GRIMSON, W. Adaptive background mixture models for real-time tracking. In *Proceedings of 1999 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'99)* (Los Alamitos, CA, USA, 1999), vol. 2, pp. 246–252. Citado na(s) página(s) 24, 25, 26, 94

[48] THIRDE, D., LI, L., AND FERRYMAN, J. Overview of the PETS 2006 Challenge. In *Proceedings of 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2006)* (Junho 2006), pp. 47–50. Citado na(s) página(s) 2, 4, 9, 26, 39, 93, 95

[49] TOYAMA, K., KRUMM, J., BRUMITT, B., AND MEYERS, B. Wallflower: principles and practice of background maintenance. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 1, pp. 255–261 vol.1. Citado na(s) página(s) 24, 25

[50] VIOLA, P., AND JONES, M. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 1* (2001), 511. Citado na(s) página(s) 15

[51] WANG, H., AND SUTER, D. A re-evaluation of mixture of gaussian background modeling. In *Proceedings of 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (2005), vol. 2, pp. 1017–1020. Citado na(s) página(s) 29, 30

[52] WELCH, G., AND BISHOP, G. An Introduction to the Kalman Filter. Tech. Rep. TR 95-041, University of North Carolina at Chapel Hill, Julho 2006. Citado na(s) página(s) 11, 82, 85

[53] WIENER, N. *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications.* The MIT Press, 1964. Citado na(s) página(s) 25

[54] WU, B., AND NEVATIA, R. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision 75*, 2 (Novembro 2007), 247–266. Citado na(s) página(s) 15

[55] YILMAZ, A., JAVED, O., AND SHAH, M. Object tracking: A survey. *ACM Computing Surveys 38*, 4 (2006). Citado na(s) página(s) 9, 10, 81