TESE DE DOUTORADO Nº 205

UMA METODOLOGIA PROBABILÍSTICA PARA COMBINAÇÃO DE DETECTORES DE PESSOAS

Natalia Cosse Batista

DATA DA DEFESA: 06/07/2015

Universidade Federal de Minas Gerais

Escola de Engenharia

Programa de Pós-Graduação em Engenharia Elétrica

UMA METODOLOGIA PROBABILÍSTICA PARA COMBINAÇÃO DE DETECTORES DE PESSOAS

Natalia Cosse Batista

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do Título de Doutor em Engenharia Elétrica.

Orientador: Prof. Guilherme Augusto Silva Pereira

Belo Horizonte - MG

Julho de 2015

Batista, Natália Cosse.
B326m Uma metodologia probabilística para combinação de detectores de pessoas [manuscrito] / Natália Cosse Batista. - 2015.
xxvi, 147 f., enc.: il.

Orientador: Guilherme Augusto Silva Pereira.

Tese (doutorado) Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 137-147.

1. Engenharia elétrica - Teses. 2. Detectores - Teses. 3. Pessoas -Teses. I. Pereira, Guilherme Augusto Silva. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 621.3(043)

"Uma Metodologia Probabilística para Combinação de Detectores de Pessoas"

Natalia Cosse Batista

Tese de Doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Doutor em Engenharia Elétrica.

Aprovada em 06 de julho de 2015.

Por:

Prof. Dr. Guilherme Augusto Silva Pereira
DEE (UFMG) - Orientador

Prof. Dr. Bruno Vilhena Adorno
DEE (UFMG)

Prof. Dr. William Robson Schwartz
DCC (UFMG)

Prof. Dr. Flávio Luis Cardeal Pádua
DECOM (CEFET-MG)

Prof. Dr. Denis Fernando Wolf
ICMC (USP)

Resumo

A coexistência de sistemas robóticos com os seres humanos é cada dia maior. Por causa disso, esses sistemas devem trabalhar de forma segura com as pessoas, sendo essencial terem uma avançada percepção do ambiente e a capacidade de detectar pessoas de forma eficiente em seu espaço de trabalho. Uma vez que a detecção de pessoas pode ser bastante desafiadora em ambientes contendo outros objetos, os detectores de pessoas existentes não são completamente confiáveis, deixando de detectar as pessoas presentes no ambiente em algumas situações, além de realizarem falsas detecções. Neste contexto, esta tese propõe uma metodologia para combinar diversos detectores de pessoas utilizando técnicas probabilísticas. O objetivo é explorar as vantagens individuais de diversos detectores de pessoas e produzir um resultado melhor em relação aos detectores individuais de forma que, quando mais de um detector indica a presença de uma pessoa em uma dada posição, a confiança da detecção é aumentada em relação à confiança de apenas um detector individual. Os detectores podem contar com informações de um ou mais sensores, tais como câmeras e sensores a laser, e sua combinação permite uma detecção de pessoas mais robusta a falhas e oclusões, resultando em informações mais precisas e completas do que aquelas fornecidas pelos detectores individuais. Além disso, por ser inspirada no Filtro de Bayes, a metodologia proposta contém etapas de predição e atualização, fazendo uso de informação temporal para melhorar a qualidade das detecções. Os modelos utilizados nessas etapas são especificados em função dos detectores combinados, levando em consideração uma medida de confiança atribuída a eles e obtida experimentalmente. Experimentos foram realizados com um robô móvel locomovendo-se autonomamente em um ambiente dinâmico contendo pessoas. A metodologia foi implementada por meio de uma grade semântica local, que representa as probabilidades da presença de pessoas em regiões específicas do ambiente. Os resultados mostraram que a metodologia proposta apresenta vantagens em relação a detectores de pessoas individuais e que é possível obter um maior número de detecções das pessoas realmente presentes no ambiente mantendo o número de falsos alarmes baixo.

Abstract

The coexistence of robotic systems and human beings is increasing every day. Because of that, these systems should be designed to interact safely with people, being essential the presence of an advanced perception of the environment and the ability to detect people efficiently. Once the detection of people can be challenging in environments containing other objects, modern people detectors are not completely reliable since they usually fail to detect people and also make false detections. In this context, this doctoral thesis proposes a methodology to combine high-level information from several people detectors using probabilistic techniques. The objective is to exploit the individual advantages of several people detectors yielding in a more accurate and complete information than the one given by a single detector alone. Thus, when more than one detector find a person at the same position, the confidence of detection is increased. The detectors rely on information from one or more sensors, such as cameras and laser rangefinders. The detectors' combination allows the prediction of the position of the persons inside the sensors' fields of view and, in some situations, outside them. Also, the fusion of the detector's output can make people detection more robust to failures and occlusions. The proposed methodology is based on a recursive Bayes Filter, whose prediction and update models are specified in function of the detectors used, taking into account a measure of confidence assigned to them, which is obtained experimentally. The concepts of prediction and update are used in the steps of the methodology, making use of temporal information to improve the quality of detection. Experiments were executed with a mobile robot that collects real data in a dynamic environment containing people while moving autonomously. The implementation of the methodology uses a local semantic grid to represent the robot's local workspace, which contains probability values related to the presence of people in specific regions of the workspace. The obtained results indicate the improvements brought by the approach in relation to a single detector alone and show that it is possible to obtain a larger number of detections of the people keeping the number of false detections low.

Agradecimentos

A Deus, por mais esta oportunidade de aprendizado, por mais um sonho realizado e por me mostrar o caminho quando as dificuldades apareceram.

Ao meu marido, por me apoiar e me incentivar sempre.

Aos meus pais, irmãs, irmãos, cunhadas, por torcerem por mim e me ajudarem de todas as formas.

À minha pequena yorkshire, por me trazer alegria.

A todas as amigas e amigos que sempre me deram força e demonstravam interesse sobre o meu trabalho.

Aos colegas do CORO e das disciplinas, por compartilharem seus conhecimentos e pela agradável companhia nos churrascos de fim de ano.

Aos professores do PPGEE, pelo aprendizado nas disciplinas e aos funcionários, sempre solícitos.

Aos professores e funcionários da UNIFEI que apoiaram o meu breve afastamento.

Aos professores da banca da minha qualificação e defesa, pelas sugestões e críticas construtivas que contribuíram para um trabalho melhor.

À FAPEMIG e à CAPES pelo apoio financeiro ao projeto.

Por fim, mas sem menos importância, ao meu orientador, por seus conselhos, sua paciência e profissionalismo.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

CORO - Laboratório de Sistemas de Computação e Robótica

FN - Falso negativo

FINT - *Fuzzy integral*

FLDA - *Fisher Linear Discriminant Analysis*

FOV - *Field Of View*

FPCPI - Falsos Positivos em relação às Células Por Imagem

FPPI - Falsos Positivos Por Imagem

GMMC - *Gaussian Mixture Model Classifier*

GPS - *Global Positioning System*

GPU - *Graphics Processing Unit*

HOG - *Histogram of Oriented Gradients*

IFR -*International Federation of Robotics*

LRF - *Local Receptive Features*

MARIA - *Manipulator Robot for Interaction and Assistance*

MCI-NN - *Minimization of interclass interference for Neural Networks*

MLP - *Multilayer perceptron*

OFM - *Occupancy Fusion Map*

PDBS - *Point Distance Based*

PDF - *Probability Density Function*

PMF - *Probability Mass Function*

RNA - Redes Neurais Artificiais

ROI - *Region Of Interest*

ROS - *Robot Operating System*

SIFT - *Scale Invariant Feature Transform*

SLAM - *Simultaneous Localization and Mapping*

SVM - *Support Vector Machines*

TFN - Taxa de Falsos Negativos

UFMG - Universidade Federal de Minas Gerais

VP - Verdadeiro positivo

VPP - Valor Preditivo Positivo

Lista de Símbolos

(a, b) - Coordenadas no espaço bidimensional do centro de uma determinada região de interesse no espaço.

α_1 **e** α_2 - Parâmetros utilizados no cálculo da variância da distribuição normal que modela o erro do movimento dos sensores relacionados à rotação.

α_3 **e** α_4 - Parâmetros utilizados no cálculo da variância da distribuição normal que modela o erro do movimento dos sensores relacionados à translação.

$atan2(c, d)$ - Função que calcula o arco-tangente de c/d cujo resultado está no intervalo $[-\pi, \pi]$.

B - Conjunto de probabilidades *a posteriori* $\{bel(x_t^{i,j})|(i, j) \in G\}$ no instante de tempo t.

B' - Conjunto de probabilidades *a priori* $\{bel(x_{t-1}^{i,j})|(i, j) \in G\}$ no instante de tempo $t - 1$.

$bel(x_t)$ - Probabilidade do estado x_t no tempo t dada a medição no instante atual z_t e a entrada de controle no instante atual u_t, ou seja, $P(x_t|z_t, u_t)$. O cálculo de $bel(x_t)$ corresponde à etapa de atualização.

$\overline{bel}(x_t)$ - Pode ser escrita como $P(x_t|z_{t-1}, u_t)$. É a probabilidade do estado x_t no tempo t dada a medição no instante passado z_{t-1} e a entrada de controle no instante atual u_t. É uma predição do estado no tempo atual baseado na probabilidade *a posteriori* do estado anterior, antes de incorporar a medição no tempo t. Na metodologia proposta, o cálculo de $\overline{bel}(x_t)$ corresponde à etapa de predição baseada no movimento dos sensores e é dada por $P(\mathrm{p}|v_t^{\text{sensores}})\overline{bel}'(x_t = \mathrm{p})$.

$\overline{bel}'(x_t)$ - De forma similar a $\overline{bel}(x_t)$, pode ser escrita como $P(x_t|z_{t-1}, u_t)$ e também é a probabilidade do estado x_t no tempo t condicionada a medições passadas $z_{1:t-1}$ e entradas de controle passados $u_{1:t}$. A diferença

entre $\overline{bel}(x_t)$ e $\overline{bel}'(x_t)$ na metodologia proposta, é que a probabilidade de uma região específica do espaço estar ocupada por pessoas é calculada com base no conhecimento do movimento das pessoas e na probabilidade de existência de pessoas na região e sua vizinhança no instante anterior $(t-1)$, ou seja, $P(x_t|z_{t-1}, v_t^{\text{pessoas}})$, que é a probabilidade do estado x_t no tempo t condicionado à medição passada z_{t-1} e a velocidade das pessoas v_t^{pessoas}. O cálculo de $\overline{bel}'(x_t)$ corresponde à etapa de predição baseada no movimento das pessoas e não leva em consideração o movimento dos sensores.

$D_1, ..., D_n$ - Conjunto de resultados de detectores de pessoas.

Δt - Intervalo de tempo entre t e $t-1$.

$dist(\cdot, \cdot)$ - Função que calcula a distância entre dois pontos em um espaço bidimensional.

η - Constante utilizada para normalização.

G - Conjunto de todas as regiões que podem ser ocupadas por pessoas.

(i, j) - Coordenadas no espaço bidimensional do centro de uma determinada região de interesse no espaço.

m - Grade de ocupação (ou mapa) que particiona o espaço em um número Q de células.

$\mathbf{m}_i$ - Célula da grade de ocupação com índice i.

μ - Média da distribuição gaussiana de velocidades nas quais uma pessoa geralmente caminha.

$\min_i$ - Função que calcula o menor valor de um conjunto com i números.

N - Número de detectores de pessoas utilizados na fusão.

np - Subconjunto do espaço amostral S contendo os eventos S - p.

p - Subconjunto do espaço amostral S contento o evento $\{X = pessoa\}$.

$p_{\textbf{imóvel}}$ - Probabilidade das pessoas permanecerem paradas em suas posições atuais.

p_k - Probabilidade calculada pelo modelo de movimento das pessoas ou dos sensores.

$p_{\mathbf{livre}}$ - Probabilidade da região (a, b) permanecer livre de pessoas.

$P(D_i = \mathbf{p}|x_t = \mathbf{p})$ - Probabilidade do detector D_i indicar a presença de pessoas quando existe pessoa na região centrada em (a, b) no tempo t.

$P(D_i = \mathbf{np}|x_t = \mathbf{p})$ - Probabilidade do detector D_i não indicar a presença de pessoas quando existe pessoa na região centrada em (a, b) no tempo t.

$P(\mathbf{m}_i)$ - Probabilidade da célula $\mathbf{m}_i$ estar ocupada.

$P(\mathbf{p})$ - Probabilidade do evento p ocorrer ou $P(X = \mathrm{p})$.

$P(\mathbf{p}|v_t^{\mathbf{sensores}})$ - Probabilidade de pessoas estarem localizadas na região centrada em (a, b) no tempo t baseada no movimento dos sensores calculada considerando a probabilidade de pessoas de outras regiões terem se movido para (a, b) em relação ao sistema de coordenadas dos sensores.

$P(\mathbf{p}|x_{t-1} = \mathbf{p}, v_t^{\mathbf{pessoas}} = 0)$ - Probabilidade das pessoas não se moverem da sua região atual (a, b).

$P(\mathbf{p}|x_{t-1} = \mathbf{np}, v_t^{\mathbf{pessoas}} = v^{\mathbf{média}})$ - Probabilidade de pessoas ocuparem a região (a, b) dado que sua velocidade é $v^{\mathrm{média}}$.

$P(z_t|x_t)$ - Probabilidade do resultado da fusão dos detectores de pessoas para uma região centrada em (a, b) no tempo t.

Q - Número de células da grade de ocupação.

$rot1$ - Rotação inicial ocorrida no movimento do robô obtida da odometria.

$\hat{rot}1$ - Rotação inicial ocorrida no movimento do robô.

$rot2$ - Rotação final ocorrida no movimento do robô obtida da odometria.

$\hat{rot}2$ - Rotação final ocorrida no movimento do robô.

s_t - Posição hipotética dos sensores no tempo t.

$\overline{s}_t$ - Posição estimada dos sensores no tempo t.

S - Espaço amostral.

σ - Desvio padrão da distribuição gaussiana de velocidades nas quais uma pessoa geralmente caminha.

t - Instante de tempo.

θ - Direção apontada pelo robô no início do movimento.

$trans$ - Translação ocorrida no movimento do robô obtida da odometria.

$\hat{trans}$ - Translação ocorrida no movimento do robô.

(v, w) - Coordenadas da célula de uma grade ou matriz bidimensional.

$v1$ - Variância da distribuição normal que modela o erro do movimento dos sensores em relação à rotação inicial.

$v2$ - Variância da distribuição normal que modela o erro do movimento dos sensores em relação à translação.

$v3$ - Variância da distribuição normal que modela o erro do movimento dos sensores em relação à rotação final.

$v^{\mathbf{angular}}$ - Velocidade angular medida do robô.

$v^{\mathbf{linear}}$ - Velocidade linear medida do robô.

$v^{\mathbf{pessoas}}$ - Velocidade das pessoas.

$v^{\mathbf{sensores}}$ - Velocidade dos sensores.

X - Variável aleatória que representa o estado da região centrada em (a, b).

x_t - Estado referente a uma determinada região do espaço centrada em (a, b) estar ocupada por pessoa ou não pessoa no instante de tempo t.

$x_t^{i,j}$ - Estado referente a uma determinada região do espaço centrada em (i, j) estar ocupada por pessoa ou não pessoa no instante de tempo t.

z_t - Resultado da fusão dos detectores de pessoas.

Lista de Algoritmos

Capítulo 1

Introdução

Nos próximos anos, o desenvolvimento de sistemas robóticos mudará drasticamente a maneira como o ser humano locomove-se, trabalha ou diverte-se. A coexistência de tais sistemas com os seres humanos aumenta a cada dia e, por causa disso, os robôs devem ser construídos para interagir de forma segura com as pessoas. Para permitir esta interação, os robôs precisam detectar pessoas de forma eficiente em seu espaço de trabalho.

Os robôs podem ser usados na execução de tarefas complexas para ajudar pessoas ou cooperar com elas, por exemplo em tarefas domésticas como a limpeza de casas, como guias em museus ou transportando cargas [Pereira et al., 2013]. Como definido pela Federação Internacional de Robótica (IFR), os robôs que realizam tais tarefas são chamados de robôs de serviço. Uma questão crítica para os robôs de serviço em um ambiente com pessoas é a interação humano-robô, pois é necessário que os robôs saibam precisamente da localização de pessoas no ambiente e, em alguns casos, é desejável que o robô seja controlado ou supervisionado por um operador humano [Ceccarelli, 2011]. Portanto, é essencial que os robôs tenham um sistema de percepção avançado, que é responsável por transformar os dados provenientes de sensores, tais como imagens de câmeras e leituras de laser, em informações consistentes e úteis para não apenas compreender o ambiente e detectar ob-

jetos, mas também para detectar pessoas e determinar sua localização. Neste contexto, um outro exemplo de aplicação são os carros autônomos, que são robôs de serviço que precisam detectar pessoas e tomar decisões críticas em tempo hábil, como evitar uma colisão, principalmente em ambientes urbanos e populosos que envolvem a possibilidade de acidentes e consequentemente risco de vida [Geronimo et al., 2010]. A detecção de pessoas também é importante em ambientes inteligentes (*Smart Environments*) em que pode ser utilizada para prever o comportamento de seus usuários [Hofmann et al., 2011].

Para os robôs móveis, tarefas como detecção de obstáculos em rota de colisão envolvem uma percepção aprimorada do entorno por meio de sensores de forma a evitar situações de risco durante a navegação do robô. Além disso, para um robô navegar com segurança, é insuficiente detectar apenas obstáculos em rota de colisão, mas também é importante diferenciar o tipo de obstáculo encontrado utilizando algoritmos, por exemplo, de Aprendizado de Máquina. Dependendo do tipo de obstáculo encontrado pelo robô, sua ação pode variar, podendo se aproximar de uma pessoa para interagir com ela, abrir uma porta ou parar completamente para uma pessoa passar. Um exemplo de robô móvel é mostrado na Figura 1.1. Esse robô têm a capacidade de locomover-se de forma autônoma e foi desenvolvido no Laboratório de Sistemas de Computação e Robótica (CORO) da UFMG, sendo utilizado em diversos projetos de pesquisa do laboratório.

1.1 Motivação

A detecção de pessoas é necessária em várias situações, sendo importante para os sistemas robóticos terem ciência da presença dessas entidades no mesmo ambiente em que estão inseridos, por diversas razões:

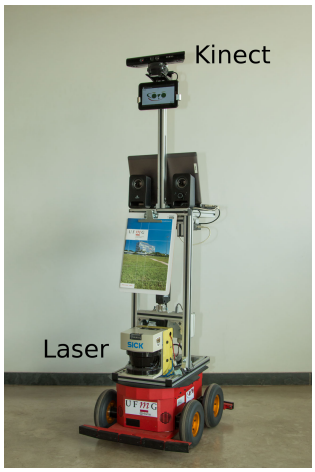


Figura 1.1: Robô móvel do Laboratório CORO/UFMG chamada MARIA (*Manipulator Robot for Interaction and Assistance*).

- Aproximação de pessoas para interação;
- Planejamento de desvios e prevenção de colisões com pessoas;
- Rastreamento de pessoas;
- Navegação em harmonia com a presença de pessoas, cedendo a preferência de passagem quando necessário;
- Redução da velocidade de um robô ao passar próximo a pessoas e locais com multidões;
- Busca e resgate de pessoas em ambientes hostis; e
- Segurança e monitoramento de ambientes.

Essa tarefa pode ser bastante desafiadora em ambientes contendo outros objetos e, além disso, a aparência das pessoas pode variar devido às suas roupas e acessórios, posição corporal, uso de dispositivos para mobilidade reduzida (como muletas, andadores e cadeiras de rodas), oclusões, dentre outros.

Para realizar a detecção de pessoas, os detectores podem utilizar dados de diversos tipos de sensores que fornecem informações que podem ser distintas e complementares sobre o ambiente. Por exemplo, um robô que possui uma câmera de infravermelho e uma câmera de visão estéreo poderá se beneficiar das diferenças tecnológicas destes sensores para perceber uma maior quantidade de informações sobre o ambiente. A câmera de infravermelho é capaz de capturar a temperatura relativa dos objetos do ambiente e pode ser usada para distinguir objetos tipicamente quentes, como pessoas e veículos, de objetos frios como árvores ou asfalto. Já a câmera de visão estéreo conta com duas câmeras de luz visível e fornece um par de imagens com o qual é possível obter a posição relativa dos objetos. Em um ambiente com boas condições de iluminação, o robô poderá detectar objetos por meio dos sensores e, além disso, obter informações sobre a distância a qual os objetos detectados estão em relação a ele. Entretanto, em ambientes com pouca iluminação (durante a noite, por exemplo) a câmera de visão estéreo terá dificuldades em obter imagens de boa qualidade e consequentemente deixará de detectar os objetos com precisão. Como o robô também possui uma câmera de infravermelho, sua capacidade de distinguir objetos, como veículos e pessoas, ainda será possível pois essa câmera tem a vantagem de detectar objetos mesmo em ambientes sem iluminação.

Esta situação ilustra uma característica da fusão de sensores, que consiste na combinação das informações de sensores distintos com o objetivo de obter informações mais completas sobre o ambiente em relação a apenas um sensor individual. Quando os dois sensores, que possuem naturezas distintas, detectam um objeto em comum, este é um forte indicativo de que o objeto realmente existe no ambiente. Porém outras possibilidades podem ocorrer, como um dos sensores detectar o objeto e o outro não e até mesmo os dois

sensores não detectarem o mesmo objeto simultaneamente, apesar do objeto estar presente no ambiente. Como os sensores estão sujeitos a ruído e possuem incertezas, as informações devem ser combinadas de forma a levar em consideração a confiança associada às detecções.

Há uma vasta literatura que trata de detecção de pessoas [Benenson et al., 2015; Geronimo et al., 2010] e seus resultados indicam que este ainda é um problema em aberto, já que os sistemas de detecção de pessoas não são completamente confiáveis e, em algumas situações, deixam de detectar pessoas presentes no ambiente além de realizarem falsas detecções. Dessa forma, o desenvolvimento de metodologias que permitam um aumento de precisão e uma redução de falsos alarmes é necessário e representa uma contribuição para área, motivando o desenvolvimento desta tese.

1.2 Objetivos

Nesta tese foi proposta uma metodologia probabilística para a fusão de detectores, com o objetivo de explorar as vantagens de diversos detectores de pessoas e produzir um resultado melhor em relação aos detectores individuais de forma que, quando mais de um detector indica a presença de uma pessoa em uma dada posição, a confiança da detecção é aumentada em relação à confiança no caso de apenas um detector detectar uma determinada pessoa. Além disso, as informações de detecções passadas são consideradas, pois pode-se prever o movimento das pessoas para saber de forma aproximada onde elas estarão no próximo instante de tempo.

Como objetivos específicos da metodologia, pode-se citar:

- Realizar a combinação de diversos detectores de pessoas para obter informações mais completas sobre as pessoas presentes no ambiente e

resultados mais confiáveis em relação à sua localização que os resultados dos detectores individuais;

- Possibilitar a detecção de pessoas que nem sempre estão no campo de visão dos sensores;
- Ser robusta a falhas dos detectores, mesmo quando todos eles deixam de detectar pessoas simultaneamente; e
- Maximizar o número de detecções de pessoas que estão presentes no ambiente mantendo o número de falsos alarmes baixo.

1.3 Contribuições

Esta tese parte da hipótese de que a combinação de diversos detectores de pessoas utilizando técnicas probabilísticas permite realizar a detecção de pessoas em ambientes não controlados de forma mais confiável e robusta em relação a detectores individuais. Em vista disso, foi proposta uma metodologia Bayesiana (Figura 1.2) que leva em consideração informações temporais e do movimento das pessoas e dos sensores, combinando-as com as informações de múltiplos detectores de pessoas, de forma a obter a probabilidade da presença de pessoas em regiões específicas do ambiente.

As contribuições da tese consistiram em vários aspectos:

- A metodologia proposta, baseada no Filtro de Bayes, possui uma fase de predição realizada em duas etapas, uma baseada no movimento do sensor e outra baseada em um modelo de movimento das pessoas;
- A proposta de um novo modelo de movimento das pessoas, que considera a probabilidade de distribuição de pessoas em regiões do espaço

e sua probabilidade de movimento. Diferentemente dos modelos de movimento existentes, o modelo de movimento das pessoas proposto é simples e genérico de forma que não é necessária a associação das pessoas na cena e a utilização de mapas do local, permitindo seu uso em ambientes desconhecidos;

- Uma descrição formal para a combinação de informações de alto nível de múltiplos detectores de pessoas na etapa de atualização, utilizando modelos baseados na precisão e na taxa de falsos negativos desses detectores, de forma a explorar as vantagens tecnológicas dos detectores; e

- A construção de uma grade de ocupação semântica com as probabilidades de regiões no espaço de trabalho local do sistema robótico estarem ocupadas por pessoas.

As contribuições citadas foram validadas por meio de experimentos com uma plataforma robótica móvel locomovendo-se autonomamente em um ambiente interno contendo pessoas. Os resultados mostraram que a metodologia proposta apresenta vantagens em relação a detectores de pessoas individuais e, em relação a outras abordagens de fusão de detectores, que geralmente au-

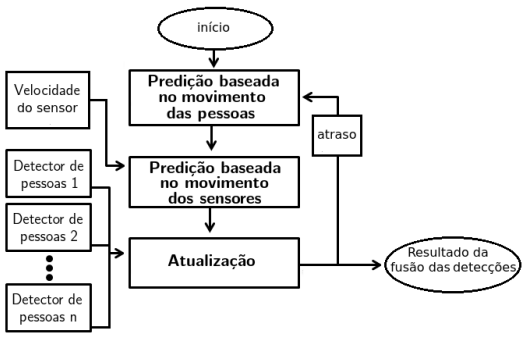


Figura 1.2: Visão geral da metodologia proposta.

mentam o número de pessoas detectadas mas também aumentam o número de falsas detecções. Os resultados desta tese demonstraram que é possível obter um maior número de detecções das pessoas realmente presentes no ambiente mantendo o número de falsos alarmes baixo.

Os resultados deste tese foram descritos nos seguintes artigos:

- Batista, N. C. e Pereira, G. A. S. (2013). Avaliação de técnicas para detecção de pedestres usando laser e câmera visando a fusão sensorial. Anais do *Simpósio Brasileiro de Automação Inteligente*.

- Batista, N. C. e Pereira, G. A. S. (2015) A Probabilistic Approach for Fusing People Detectors. *Journal of Control, Automation and Electrical Systems*, 26(6): 616-629.

1.4 Organização da Tese

O próximo capítulo apresenta os conceitos teóricos relacionados à metodologia proposta e à sua implementação. O estado da arte envolvendo técnicas de detecção de pedestres é descrito no Capítulo 3. A metodologia proposta para a fusão de detectores de pessoas é apresentada no Capítulo 4, bem como a descrição de uma implementação utilizando grades de ocupação semânticas. No Capítulo 5 são descritos os experimentos realizados em bases de dados reais e é realizada a análise dos resultados. Finalmente, as considerações finais sobre a tese e os trabalhos futuros são abordados no Capítulo 6.

Capítulo 2

Conceitos preliminares

Este capítulo apresenta uma breve introdução aos conceitos teóricos relacionados à metodologia proposta e à sua implementação. Por se tratar de uma abordagem com foco em detecção de pessoas, são apresentados os conceitos de detectores de pessoas e seu funcionamento. Existem detectores de pessoas baseados em diversos tipos de sensores (laser, câmera, etc), cada um com suas vantagens mas também com limitações. Para maximizar o resultado da detecção de pessoas, a proposta desta tese é combinar diversos detectores e, além disso, utilizar informações temporais por meio de uma abordagem probabilística. Em seguida, como a metodologia proposta é inspirada no Filtro de Bayes, este é descrito na Seção 2.2. A metodologia tem a vantagem de ser flexível em vários aspectos, por exemplo: 1) na utilização de um número variável de detectores associados a diversos sensores; 2) em suas aplicações; pois parte do pressuposto que o ambiente no qual o robô está inserido é desconhecido; e 3) em sua implementação prática, que permite a utilização de mais de uma técnica. Na sequência, as técnicas de implementação são descritas (Seção 2.3), com enfoque nas grades de ocupação semânticas, que é a técnica selecionada para a implementação utilizada nos experimentos.

2.1 Detecção de pessoas

A detecção de pessoas é uma tarefa que consiste em encontrar automaticamente pessoas em determinada região do espaço por meio do processamento de dados de sensores. Essa tarefa, que pode ser executada em ambientes complexos com vários objetos, envolve a classificação e a localização das pessoas.

A classificação de objetos está inserida no campo de Aprendizado de Máquina e também é conhecida como reconhecimento de padrões. De acordo com Nilsson [1998], Aprendizado de Máquina refere-se a alterações na estrutura, programas ou dados de sistemas que realizam tarefas tais como reconhecimento de padrões, diagnósticos, planejamento, controle de robôs, predição, etc, visando melhorar seu desempenho. O objetivo do Aprendizado de Máquina é fazer com que máquinas aprendam ou descubram automaticamente determinados padrões ou comportamentos a partir de exemplos ou observações de dados por meio do uso de algoritmos computacionais. Há um vasto campo de aplicações para a área de Aprendizado de Máquina, por exemplo reconhecimento de objetos ou voz, extração de relacionamento de dados em grandes bases, previsão do comportamento de clientes a partir de dados de compras, sistemas de suporte ao diagnóstico de doenças, detecção de fraudes bancárias e máquinas de busca na web.

Os detectores podem utilizar as observações de um ou mais sensores, realizando a classificação dessas observações pela atribuição de determinadas categorias ou classes a elas [Theodoridis e Koutroumbas, 2008; Devroye et al., 1996]. Neste contexto, uma observação é um conjunto de medições numéricas tais como uma imagem, um vetor de distâncias dos objetos ou um vetor de temperaturas. A classificação das observações resume-se a criar uma função ou um mapeamento que relaciona cada observação a uma classe. Para a

detecção de pessoas, a classificação de objetos deve ter uma das classes sendo a classe de pessoas. A localização dos objetos classificados como pessoas no ambiente pode ser aproximada por meio de transformações de coordenadas e mapeamentos que dependem do tipo de sensor utilizado. Há sensores, como o sensor de distância a laser, que permitem uma aproximação mais precisa, pois já fornecem as distâncias e ângulos dos objetos. Outros sensores, como câmeras, não fornecem diretamente a distância, que deve ser estimada por meio de mapeamentos de acordo a natureza do sensor. Por exemplo, com uma câmera de visão estéreo é possível obter a posição relativa dos objetos utilizando os pixeis correspondentes do par de imagens e a geometria da câmera [Saxena et al., 2007].

A subseção a seguir descreve os tipos de sensores mais utilizados para a detecção de pessoas.

2.1.1 Sensores

Vários tipos de sensores são utilizados para a detecção de pessoas, tais como o sensor de distância a laser, radar, câmera para luz visível e câmera para infravermelho. Esses sensores podem ser divididos em duas categorias: passivos e ativos. Conforme Bertozzi et al. [2002], a utilização de sensores passivos, como as câmeras, apresenta vantagens em relação aos sensores ativos (laser, sonar e radar). As câmeras apresentam a possibilidade de adquirir dados de uma maneira não invasiva, ou seja, sem alterar o ambiente, e possuem informações visuais que são ricas em detalhes e possibilitam o reconhecimento de objetos e a classificação de texturas [Discant et al., 2007]. Em tarefas como a detecção de pessoas, conforme Broggi et al. [2009], a visão é a melhor tecnologia para detectar a presença de pessoas parcialmente oclusas. Outra vantagem é o preço de aquisição mais baixo. A desvantagem

dos sensores baseados em visão é que são menos robustos a neblina, falta de luminosidade ou condições de luz solar direta e não são adequados para medir longas distâncias com precisão.

Por outro lado, os sensores ativos têm a vantagem de poderem medir movimento de uma forma mais direta e com menor quantidade de dados que os sensores baseados em visão, resultando em um processamento mais rápido. Este tipo de sensor pode detectar obstáculos com grande acurácia mesmo na escuridão, pois é robusto às condições de iluminação [Antunes et al., 2012]. Um laser pode operar a longas distâncias além da possibilidade de medir a distância de objetos com maior precisão que os sensores passivos. Entretanto, podem sofrer interferência de outros sensores do mesmo tipo, o que prejudica sua confiabilidade no uso em massa. Outra desvantagem é que o laser não provê cor nem textura dos objetos, dificultando a distinção entre eles. Outra situação em que apresenta desvantagem é na tarefa de detecção de pessoas andando em grupo e muito próximas, pois os dados de uma pessoa podem se fundir aos de outra.

A tendência atual é realizar a fusão de múltiplos sensores em diversas aplicações da robótica, pois a fusão permite combinar as vantagens tecnológicas dos sensores para compensar suas limitações. Além disso, a redundância de informação que pode ser obtida com a fusão deve ser explorada para maximizar a confiabilidade da percepção ou para reduzir o espaço de busca em outros sensores, consequentemente diminuindo o tempo de processamento.

Em trabalhos de fusão sensorial, encontram-se vários tipos de sensores combinados sendo os mais comuns as câmeras de visão monoculares e de visão estéreo, sensores de infravermelho e sensores a laser [Premebida e Nunes, 2013; Geronimo et al., 2010; Teixeira et al., 2010]. A câmera de visão monocular, que possui um sensor de imagem digital, capta a luminosidade da cena

que é projetada sobre ele, dentro espectro visível, gerando uma imagem digital bidimensional formada por uma matriz de pontos, onde cada ponto é um pixel, que codifica a intensidade da luz em determinada região do ambiente. Uma imagem pode ser definida como uma função bidimensional, f(x,y), em que x e y são coordenadas espaciais no plano e a amplitude de f em qualquer (x,y) é chamada de intensidade ou nível de cinza [Gonzalez e Woods, 2003].

As câmeras que obtêm um par estéreo de imagens (Figura 2.1) podem ser utilizadas para estimar distância dos objetos realizando-se uma reconstrução 3D por meio da associação de pixeis das duas imagens e triangulação [Antunes et al., 2012], por exemplo, obtendo um mapa de disparidade. Entretanto, o alcance é limitado, como relatado em [Lima e Pereira, 2010], pois a variação não linear da disparidade com a profundidade gera problemas de valores não contínuos e não confiáveis para grandes distâncias. Além da distância ser limitada, as câmeras de par estéreo são sujeitas a erros em superfícies inclinadas com textura homogênea, conforme mostrado em [Antunes et al., 2012].

Os sensores de infravermelho capturam a temperatura relativa dos objetos da cena e podem ser usados para distinguir objetos tipicamente quentes como pessoas e veículos de objetos frios como árvores ou asfalto. A vantagem desse sensor é que podem detectar pessoas mesmo em ambientes sem iluminação [Elguebaly e Bouguila, 2011; Ge et al., 2009; Linzmeier et al., 2004].

O sensor de distância a laser, que é um sensor de proximidade que opera

Figura 2.1: Câmera Bumblebee2 da Point Grey Research (http://ww2.ptgrey.com/stereo-vision/bumblebee-2).

medindo o tempo de ida e volta de pulsos de luz, apresenta campo de visão amplo e frequências altas de leitura. É utilizado em vários trabalhos da literatura para detectar pessoas [Weinrich et al., 2014; Mozos et al., 2010; Premebida et al., 2009; Spinello e Siegwart, 2008], além disso possui alcance maior que das câmeras e é mais confiável para medir distâncias. Nestes trabalhos nota-se uma maior utilização de lasers de uma camada ou múltiplas camadas e a pouca utilização de lasers 3D. Esse tipo de laser possui desvantagens em relação aos lasers 2D devido à sua complexidade, tamanho volumoso e baixas taxas de amostragem [Zhi-yu et al., 2001]. Além disso mostra-se inviável em aplicações comerciais devido ao seu alto custo. Por exemplo, o laser SICK LMS-291-S05 (Figura 2.2), com taxa de amostragem de 75 Hz, custa aproximadamente 4.000 dólares, enquanto que o laser 3D Velodyne HDL-64E tem taxa de amostragem entre 5 e 15 Hz e valor próximo de 75.000 dólares.

Outros sensores de proximidade também utilizados em trabalhos de detecção de pessoas é o sonar (sensor de ultrassom) e o radar (sensor de ondas de rádio). O sonar provê as distâncias dos objetos localizados dentro dos limites de seu cone de radiação, porém com grande incerteza em relação aos ângulos. Esse sensor está sujeito a diversos erros de medição, por exemplo devido a múltiplas reflexões, variações de temperatura, umidade e pressão [Bu e Chan, 2005; Ribo e Pinz, 2001]. O radar emite uma onda eletromagnética e, com base na análise dos sinais que retornam à antena, a distância e velocidade dos objetos podem ser calculadas com boa precisão. Além disso,

Figura 2.2: Laser SICK LMS-291-S05 (http://www.sick.com).

o radar pode operar em condições ambientais diversas (neblina, neve, etc) [Teixeira et al., 2010; Bu e Chan, 2005].

Os sensores obtém observações do ambiente que serão posteriormente utilizadas na detecção de pessoas. Para isso, essas observações devem ser processadas de forma a extrair características ou atributos que as representem. As características, explicadas na próxima subseção, são então analisadas por um classificador que determinará a classe a que pertence a observação.

2.1.2 Características

As pessoas apresentam certas feições ou peculiaridades que as diferenciam de outros objetos e permitem sua detecção. Algumas dessas peculiaridades são: a forma, que pode ser detectada por câmeras, radar e laser; o movimento típico ao caminhar ou correr, que pode ser detectado por câmeras, sonar, laser e radar; e até o movimento involuntário de órgãos do corpo como coração e pulmão, que pode ser medido com ondas de radio que penetram a pele e sinais de ultrassom [Teixeira et al., 2010]. As características são medições que descrevem os objetos e possibilitam quantificar as suas múltiplas feições.

As características são utilizadas para extrair informações discriminativas que reduzam a dimensionalidade do conjunto de amostras e são expressas por meio de determinadas medidas compondo um espaço com dimensionalidade igual ao número de medidas, denominado espaço de características [Pedrini e Schwartz, 2008]. As características devem ser escolhidas de forma que amostras da mesma classe tenham características espacialmente próximas e amostras de classes distintas tenham características distantes, ou seja, boa capacidade discriminatória.

Pode-se citar como características o número de pontos, linhas, quinas, bordas (tamanho, regularidade, curvatura, etc), cor, intensidade, contraste,

textura, forma e localização espacial, dentre outras. Para melhor compreensão dos trabalhos da literatura relacionados à detecção de pessoas, serão explicados resumidamente os seguintes descritores:

- Histograma: um histograma é uma representação estatística das frequências das observações em determinado intervalo. É composto por intervalos discretos (*bins*) que contém a frequência de uma determinada classe. Por exemplo, o histograma de uma imagem digital com níveis de intensidade no intervalo $[0, L-1]$, é uma função discreta $h(r_k) = n_k$, onde r_k é o k-ésimo valor de intensidade e n_k é o número de pixeis da imagem com intensidade r_k [Gonzalez e Woods, 2003].

- Histograma de orientação de bordas: é um histograma que realiza a contagem das direções dos gradientes das bordas [Geronimo e Lopez, 2014]. A borda de um subconjunto de pixeis conexos S de uma imagem é o conjunto de pixeis pertencentes a S que possuem vizinhança-4 com um ou mais pixeis externos a S [Pedrini e Schwartz, 2008]. Os ângulos das bordas podem ser detectados por operadores como o de Sobel. A Figura 2.3 ilustra o conceito de bordas e seu histograma de orientação.

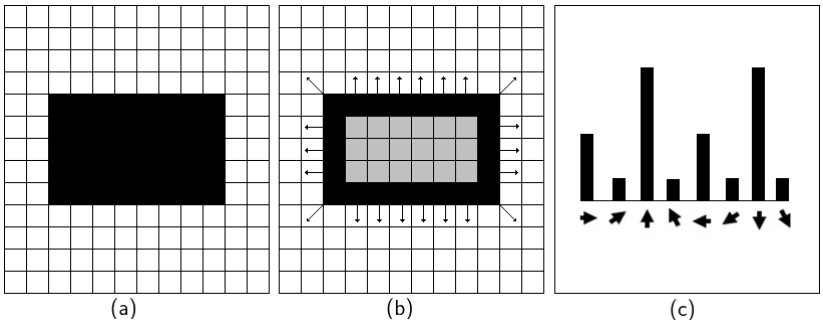


Figura 2.3: Ilustração do conceito de bordas e seu histograma de orientação. (a) Subconjunto de pixeis conexos em preto. (b) Pixeis da borda em preto e pixeis do interior em cinza. (c) Histograma ilustrativo da orientação das bordas.

- Características de Haar: as características baseadas em filtros de Haar, que consistem em cálculos de médias e diferenças entre pixeis [Pedrini e Schwartz, 2008], foram desenvolvidas por Viola e Jones [2001] e consideram regiões retangulares adjacentes em uma localização específica na janela de detecção da imagem. A Figura 2.4 mostra exemplo dessas características retangulares, relativas à janela de detecção. A soma dos pixeis dentro dos retângulos brancos é subtraída da soma dos pixeis nos retângulos pretos. Cada tipo de característica pode indicar a presença de certas propriedades na imagem, tais como bordas ou mudanças na textura.

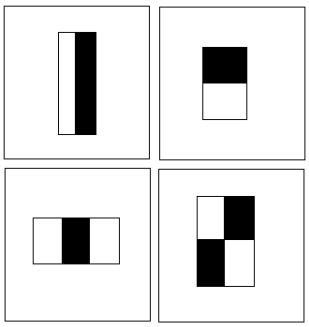


Figura 2.4: Exemplos de características de Haar. Imagem baseada na Figura 1 de Viola e Jones [2001].

- HOG: os histogramas de gradientes orientados (HOG - *Histogram of Oriented Gradients*) são similares aos histogramas de orientação de bordas, SIFT e descritores de forma, mas são calculados em uma grade densa com células uniformemente espaçadas. A imagem é analisada em múltiplas escalas por meio de janelas, que são divididas em regiões pequenas (as células), cada uma tendo um histograma local de direções dos gradientes. Os gradientes são computados utilizando suavização gaussiana seguida de várias máscaras de derivações discretas. As células são então agrupadas em regiões chamadas blocos. Os blocos individuais

são normalizados para reduzir a influência do brilho da imagem e são usados para obter o descritor [Dalal e Triggs, 2005].

2.1.3 Classificadores

Os classificadores são algoritmos para rotular ou classificar objetos dentro de um conjunto discreto de classes e fazem parte da área de Aprendizado de Máquina, mais especificamente do aprendizado supervisionado [Alpaydin, 2010]. No aprendizado supervisionado é necessário um conjunto de dados de treinamento composto pelas características das amostras de exemplo e seus respectivos rótulos (Figura 2.5). Neste tipo de aprendizado incluem-se algoritmos de classificação e de regressão. Já no aprendizado não-supervisionado, os dados de treinamento consistem em um conjunto de amostras sem o conhecimento dos rótulos correspondentes. Métodos de aprendizado não-supervisionados são utilizados para agrupar dados dos quais não se conhece o rótulo, sendo conhecidos também como métodos de agrupamento (*clustering*), em que os rótulos associados a cada categoria são obtidos apenas por meio dos dados. Um exemplo de algoritmo de agrupamento de simples implementação é o *k-means*, que particiona um conjunto de pontos entre k subconjuntos disjuntos de forma a minimizar a distância intra-classe e maximizar a distância inter-classe.

A classificação tem o objetivo de associar o rótulo de uma categoria ou classe (saída) a uma amostra de entrada, como definido em Lorena e de Carvalho [2007]:

> Dado um conjunto de exemplos rotulados na forma $(x_i; y_i)$, em que x_i representa um exemplo e y_i denota o seu rótulo, deve-se produzir um classificador, também denominado modelo, preditor ou hipótese, capaz de predizer precisamente o rótulo de novos

dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função f, a qual recebe um dado x e fornece uma predição y.

Se a saída desejada consistir em uma ou mais variáveis contínuas, o problema é chamado de regressão [Bishop, 2006].

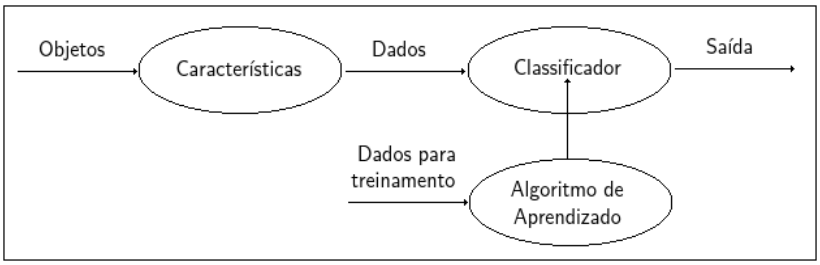


Figura 2.5: Aprendizado Supervisionado.

A seguir serão apresentados sucintamente algoritmos de classificação utilizados em detecção de pessoas:

1. **AdaBoost**: o AdaBoost foi o primeiro algoritmo prático de Boosting que é um método para criar um classificador forte (ou de melhor acurácia) pela combinação de um conjunto de classificadores fracos, cuja acurácia é melhor do que escolhas aleatórias [Spinello e Siegwart, 2008].

   Conforme explicado no trabalho de Schapire [2013], dado um conjunto de m amostras de treinamento em determinado domínio D e seus respectivos rótulos, calcula-se uma distribuição D_t, onde $t = 1, \ldots, T$ é a iteração corrente, utilizada para treinar um classificador fraco e obter uma hipótese (ou classificador) $h_t : D \rightarrow \{-1, +1\}$. O objetivo do classificador fraco é encontrar uma hipótese fraca cujo erro seja pequeno em relação à D_t. A hipótese final H calcula o sinal de uma combinação ponderada de hipóteses fracas:

$$F(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}), \tag{2.1}$$

onde $\mathbf{x}$ é o vetor de características representando uma amostra. De acordo com a essa equação, pode-se dizer que H é calculado por uma votação ponderada das hipóteses fracas h_t associadas a um peso α_t.

2. **Aprendizado Bayesiano**: dado um conjunto de amostras e seus respectivos rótulos para treinamento, o classificador Bayesiano utiliza a probabilidade *a posteriori* $P(w_i|\mathbf{x})$ de um determinado vetor de características $\mathbf{x}$ de uma amostra apresentada pertencer a uma classe w_i, para $i = 1, 2, ..., m$, onde m é o número de classes, para seguir a regra de decisão [Pedrini e Schwartz, 2008]:

$$P(w_i|\mathbf{x}) > P(w_j|\mathbf{x}),\ j = 1, \ldots, m; i \neq j. \tag{2.2}$$

Essa regra atribui a um vetor de características $\mathbf{x}$ um rótulo de modo que cada amostra pertença à classe que maximize a probabilidade *a posteriori*.

Utilizando o Teorema de Bayes (Seção 2.2), é possível particionar o espaço de características em regiões que representam as classes por meio da regra de Bayes para taxa mínima de erro, possibilitando atribuir um rótulo a $\mathbf{x}$ dependendo da sua localização no espaço de características. O Teorema de Bayes é usado para converter a probabilidade *a priori* $P(w_i)$ em probabilidade *a posteriori* incorporando a evidência provida pela observação.

3. **Naive Bayes**: semelhante à classificação Bayesiana, porém as proba-

bilidades são calculadas de forma distinta. Além disso, assume-se a independência condicional entre os elementos do vetor de características. Cada característica tem uma PDF (Função Densidade de Probabilidade) modelada por uma Gaussiana e, dado o vetor de características $\mathbf{x}$, atribui-se a amostra à classe w_i que apresenta a maior probabilidade condicional $P(w_i|\mathbf{x})$ [Pedrini e Schwartz, 2008].

4. **Redes Neurais**: Redes Neurais Artificiais (RNA) são sistemas compostos por neurônios artificiais interligados, representados por funções matemáticas que mapeiam um espaço de entrada a um espaço de saída. Cada conexão possui um determinado peso e os neurônios podem estar organizados em uma ou mais camadas em uma rede. As redes neurais têm a capacidade de aprender sobre o seu ambiente por meio de algoritmos de aprendizagem para ajuste dos pesos dos neurônios na etapa de treinamento. Como exemplos de modelos de redes neurais pode-se citar o perceptron de uma camada, as MLPs (*Multilayer Perceptron*) de múltiplas camadas, as redes de função de base radial, dentre outras [Haykin, 2001]. As RNAs podem ser usadas para resolver diversos tipos de problemas, por exemplo de classificação de padrões, devido à sua capacidade de aprender por meio de exemplos e de generalizar, ou seja, oferecer respostas consistentes para dados não apresentados durante a etapa de treinamento [Braga et al., 2007].

5. **MCI-NN**: consiste na minimização de interferência inter-classe (MCI - *Minimization of interclass interference*), que é um algoritmo de treinamento para redes neurais baseado em margem máxima. O MCI cria uma camada de saída escondida na RNA na qual os padrões possuem a distribuição estatística pretendida, por meio da substituição da ca-

mada de saída linear pelo kernel de Mahalanobis, que leva em conta as correlações do conjunto de dados [Premebida et al., 2009].

6. **FLDA**: o FLDA (*Fisher Linear Discriminant Analysis*) é um método para encontrar uma combinação linear de características que possibilita separar duas ou mais classes de objetos. O problema de otimização do FLDA consiste em maximizar a razão entre a variância entre as classes e a variância intra-classe, utilizando o resultado para encontrar o plano que melhor separa as classes [Premebida et al., 2009].

7. **GMMC**: para o classificador GMMC (*Gaussian Mixture Model Classifier*), utiliza-se uma distribuição Gaussiana multivariada para cada amostra. A probabilidade de uma amostra pertencer a uma classe pode ser calculada considerando uma composição linear de PDFs Gaussianas, definidas por um vetor de pesos, um vetor da média e uma matriz de covariância [Premebida et al., 2009].

8. **Máquinas de Vetores de Suporte**: o objetivo das Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) é a construção de hiperplanos que separam os exemplos positivos e negativos de uma classe com a maior margem possível, que é uma medida de confiança da previsão do classificador. Seja T um conjunto de treinamento com n dados $x_i \in X$ e seus respectivos rótulos $y_i \in Y$, em que X constitui o espaço dos dados e $Y = \{-1, 1\}$. T é linearmente separável se é possível separar os dados das classes $+1$ e -1 por um hiperplano [Lorena e de Carvalho, 2007]. Os dados que se encontram mais próximos ou sobre os hiperplanos são denominados vetores de suporte e são considerados os dados mais informativos do conjunto de treinamento, pois somente eles participam na determinação da equação do hiper-

plano separador. Como resultado final, obtém-se uma função linear otimizada com a melhor capacidade de generalização, que representa o hiperplano separador o qual divide os dados com a maior margem. Em aplicações cujos dados não sejam linearmente separáveis, o conjunto de treinamento é mapeado de seu espaço original para um novo espaço de maior dimensão, denominado espaço de características, por meio de uma função denominada *kernel* [Burges, 1998].

2.2 Filtro de Bayes

Seja X uma variável aleatória e x um valor específico que X possa assumir. De forma a simplificar a notação, será utilizada a mesma notação de Thrun et al. [2005] para a probabilidade $P(X = x)$ de X assumir o valor x, que será escrita como $P(x)$.

O Teorema de Bayes é baseado na probabilidade condicional e relaciona a probabilidade condicional $P(x|z)$ a $P(z|x)$:

$$P(x|z) = \frac{P(z|x)P(x)}{P(z)}. \tag{2.3}$$

Este teorema pode ser calculado utilizando a Lei da Probabilidade Total. No caso discreto, dado que $P(z) > 0$, a probabilidade $P(x|z)$ é calculada por:

$$P(x|z) = \frac{P(z|x)P(x)}{\sum_{x'} P(z|x')P(x')}. \tag{2.4}$$

Na área de robótica probabilística, o Teorema de Bayes é utilizado para inferir uma quantidade x a partir de dados z provenientes de sensores. Neste caso, x denota um estado que pode ser, por exemplo, a localização e orientação de um robô (pose), a velocidade de um robô, a localização e características de objetos no ambiente, etc. A distribuição $P(x)$ é conhecida como probabili-

dade *a priori* e representa o conhecimento a respeito de x antes de incorporar os dados z. A probabilidade $P(x|z)$ é conhecida como probabilidade *a posteriori* (após o experimento ser conduzido) e $P(z|x)$ pode ser considerada um modelo generativo, pois descreve como o estado x causa a medição dos sensores z [Thrun et al., 2005].

O Filtro de Bayes é derivado da aplicação do teorema de Bayes à probabilidade *a posteriori* $P(x_t|z_{1:t}, u_{1:t})$, que condiciona o estado x_t à medição de sensores $z_{1:t}$ e a informações de controle $u_{1:t}$. Como os estados podem mudar ao longo do tempo e medições e controle são realizados em determinados instantes de tempo, o subscrito nos valores das variáveis aleatórias indicam o instante de tempo considerado, em valores discretos. Por exemplo, $z_{1:t}$ representa o conjunto de todas as medições adquiridas do instante de tempo igual a 1 até o instante t. Essa notação é equivalente à utilizada em Thrun et al. [2005], assim como a denominação do conhecimento de determinado estado ou crença do estado (*belief*):

- $bel(x_t) = P(x_t|z_{1:t}, u_{1:t})$: probabilidade do estado x_t no tempo t condicionada a todas as medições passadas $z_{1:t}$ e entradas de controle passados $u_{1:t}$. As medições são usadas para corrigir a predição do estado, incorporando os novos dados medidos neste instante.

- $\overline{bel}(x_t) = P(x_t|z_{1:t-1}, u_{1:t})$: probabilidade do estado x_t no tempo t condicionada a medições passadas $z_{1:t-1}$ e entradas de controle passados $u_{1:t}$. É uma predição do estado no tempo atual baseado na probabilidade *a posteriori* do estado anterior, antes de incorporar a medição no tempo t.

A derivação do Filtro de Bayes apresentada em Thrun et al. [2005] assume que o estado é completo, ou seja, as condições de Markov são obedecidas.

Considera-se que um estado é completo se este é o melhor preditor do futuro, ou seja, se o conhecimento sobre estados passados, medições e controle não contribuem com informações adicionais que possam melhorar a predição. Nessas condições, estados passados e futuros são independentes, porém, em situações práticas, essas condições são violadas por erros de modelagem e de aproximação, dentre outros fatores. Entretanto, o Filtro de Bayes é robusto a violações dessas condições. Dessa forma, o subscrito nos valores das variáveis aleatórias que indica o conjunto de todas as medições adquiridas nos instantes passados pode ser suprimido.

As equações do Filtro de Bayes, dadas a probabilidade $bel(x_{t-1})$, a entrada de controle mais recente u_t e a medição mais recente z_t, são [Thrun et al., 2005]:

$$\overline{bel}(x_t) = \int P(x_t|u_t, x_{t-1})bel(x_{t-1})dx_{t-1} \tag{2.5}$$

$$bel(x_t) = \eta P(z_t|x_t)\overline{bel}(x_t). \tag{2.6}$$

A constante η é utilizada para normalização, garantindo que o produto resultante seja uma função de probabilidade e que sua integral seja igual a 1. O Filtro de Bayes é recursivo e calcula a probabilidade $bel(x_t)$ no tempo t utilizando a probabilidade $bel(x_{t-1})$ no tempo $t-1$. A primeira equação leva em consideração o controle u_t e probabilidade do estado x_{t-1} para calcular a probabilidade do estado x_t, etapa que é conhecida como predição. A segunda equação é conhecida como atualização e leva em consideração a medição z_t. Quando o espaço de estados é finito, a integral da primeira equação torna-se um somatório finito.

Na formalização da metodologia proposta nesta tese (Capítulo 4), serão utilizadas três equações baseadas no Filtro de Bayes: a primeira para calcular $\overline{bel}'(x_t)$ na primeira etapa da predição, a segunda para calcular $\overline{bel}(x_t)$ na

segunda etapa da predição e a terceira para calcular $bel(x_t)$, na etapa de atualização.

2.3 Grades de ocupação semânticas

Em aplicações práticas, os filtros gaussianos (tais como o Filtro de Kalman e suas variações) e os filtros não paramétricos (tais como o filtro de partículas) são implementações tratáveis do Filtro de Bayes para espaços contínuos. As abordagens paramétricas calculam a probabilidade *a posteriori* por meio de uma distribuição conhecida com um número fixo de parâmetros, por exemplo a distribuição normal (ou gaussiana), que é parametrizada por sua média e desvio padrão. Como exemplo pode-se citar o Filtro de Kalman, que implementa o cálculo da crença para estados contínuos e assume que as probabilidades *a priori* são gaussianas. O Filtro de Kalman é utilizado para filtragem e predição de sistemas lineares e, na aplicação de detecção de pessoas, pode ser utilizado para rastreamento [Bertozzi et al., 2004; Schneider e Gavrila, 2013], sendo robusto a problemas relacionados a ruído dos sensores e associação temporal. Entretanto, não provê informação sobre a probabilidade de existir pessoas por regiões, pois o foco é na posição das pessoas que foram detectadas pelo menos uma vez. As abordagens não paramétricas aproximam a probabilidade *a posteriori* por um número finito de valores correspondendo a partes do espaço de estados, não tendo diretamente um modelo em forma de função nem parâmetros fixos (média, variância, etc). Porém, seu comportamento pode ser caracterizado por uma representação gráfica [Aguirre, 2007; Thrun et al., 2005]. Como exemplos de abordagens não paramétricas pode-se citar o Filtro de Partículas, histogramas e grades de ocupação bayesianas.

Neste trabalho de doutorado, são utilizadas as grades de ocupação semânticas, derivadas das grades de ocupação, como opção para a implementação da metodologia proposta, que é baseada em um Filtro Bayesiano. Uma grade de ocupação é uma representação estocástica da informação espacial do ambiente [Elfes, 1990], sendo uma abordagem não paramétrica. A grade ou matriz bidimensional é arranjada em células de mesmo tamanho, cada uma associada a uma coordenada (v, w) e à sua probabilidade de ocupação. Como definido em [Thrun et al., 2005], seja $\mathbf{m}_i$ a célula com índice i. A grade de ocupação (ou mapa) m particiona o espaço em um número Q de células:

$$m = \{\mathbf{m}_i | 1 \leq i \leq Q\}, \tag{2.7}$$

Cada célula $\mathbf{m}_i$ possui um valor de ocupação, tradicionalmente binário, para indicar se a célula está ocupada (valor igual a 1) ou livre (valor igual a 0). A probabilidade da célula estar ocupada é referida como $P(\mathbf{m}_i)$. Considerando independência entre as células, cada probabilidade pode ser calculada separadamente por célula, considerando o conjunto de medições do sensor $z_{1:t}$ e a sequência de posições do sensor $x_{1:t}$ até o tempo t:

$$P(\mathbf{m}_i | z_{1:t}, x_{1:t}) \tag{2.8}$$

Utilizando uma formulação Bayesiana e considerando que $P(\mathbf{m}_i)$ independe da posição do sensor no ambiente, a probabilidade da célula estar ocupada pode ser aproximada por [Thrun, 2003]:

$$P(\mathbf{m}_i | z_{1:t}) \;=\; \frac{P(z_t | \mathbf{m}_i) P(\mathbf{m}_i | z_{1:t-1})}{P(z_t | z_{1:t-1})}, \tag{2.9}$$

onde $P(z_t|\mathbf{m}_i)$ é o modelo do sensor. Aplicando-se a regra de Bayes ao termo $P(z_t|\mathbf{m}_i)$, obtém-se:

$$P(\mathbf{m}_i|z_{1:t}) = \frac{P(\mathbf{m}_i|z_t)P(z_t)P(\mathbf{m}_i|z_{1:t-1})}{P(\mathbf{m}_i)P(z_t|z_{1:t-1})}. \quad (2.10)$$

A grade de ocupação é a representação de um ambiente com a lista de objetos presentes e suas posições [Thrun et al., 2005]. As grades são utilizadas em várias tarefas, tais como localização de robôs, planejamento de rotas, desvio de obstáculos e rastreamento de objetos, tendo também diversos algoritmos para sua geração. Por exemplo, o rastreamento de objetos utilizando o Filtro de Ocupação Bayesiano, que utiliza grades de ocupação. Além de fornecer as informações de ocupação, possibilita o rastreamento das células associando-as também a uma velocidade [Yoder et al., 2014; Ros e Mekhnacha, 2009; Tay et al., 2008].

Em geral, os trabalhos de mapeamento são concentrados em ambientes estáticos, podendo haver ou não a associação de conceitos semânticos ao mapa. Os conceitos semânticos ajudam a explicar funcionalidades do ambiente, sua estrutura e conectividade, utilizando conceitos como cômodos, corredores, portas, etc [Wang e Chen, 2011]. De acordo com Wolf e Sukhatme [2008], o problema do mapeamento semântico consiste em utilizar sensores móveis para criar mapas que representem não apenas a ocupação mas também outras propriedades do ambiente, atribuindo significado aos espaços. Esses mapas semânticos podem armazenar informação sobre objetos, funcionalidades ou eventos do ambiente [Biresev, 2012].

As informações semânticas podem colaborar para uma melhor percepção do ambiente e, em aplicações de robótica, auxiliar os robôs no planejamento de seu comportamento no ambiente. Uma grade de ocupação semântica integra a representação espacial do ambiente com as localizações dos objetos de classes conhecidas [Nüchter e Hertzberg, 2008]. Embora seja conhecida

como uma técnica que demanda considerável tempo de processamento, a utilização de grades de ocupação apresenta vantagens como a facilidade de fusão de vários sensores e a integração da detecção de outros objetos, além de fornecer informação sobre a ocupação do espaço.

Dadas as definições e conceitos apresentados neste capítulo, no próximo capítulo são apresentados diversos trabalhos cujo objetivo é detectar a presença de pessoas no ambiente.

Capítulo 3

Estado da arte

Neste capítulo são descritos os trabalhos relacionados à metodologia proposta na tese. Há uma vasta literatura que trata de detecção de pessoas. Os sistemas para detecção de pessoas geralmente buscam candidatos dentro do campo de visão dos sensores usando características como forma, simetria, textura, movimento e periodicidade do movimento de pernas humanas [Broggi et al., 2009]. Embora a metodologia proposta possibilite a utilização de vários detectores de pessoas associados a diversos sensores, dois sensores serão destacados nas três seções a seguir: o sensor de distância a laser e a câmera de luz visível. Estes sensores possuem vantagens conhecidas (Capítulo 2) e foram escolhidos para os experimentos desta tese (Capítulo 5) pois fornecem informações complementares e de naturezas distintas. Uma forte tendência observada nos trabalhos de detecção de pessoas é a fusão sensorial, que mostrou-se fundamental para a robustez da abordagem nos experimentos realizados, pois os sensores individualmente deixam de detectar pessoas em determinadas situações (oclusões, más condições de iluminação, etc). Assim, após detalhar os trabalhos de detecção de pessoas com laser e com câmera, neste capítulo são apresentados trabalhos sobre fusão sensorial.

A metodologia proposta envolve também o Filtro de Bayes, amplamente utilizado em diversas aplicações por permitir integrar informações passadas

com as observações atuais dos sensores para predizer o estado atual. As abordagens bayesianas são detalhadas na Seção 3.4. Para concluir a revisão do estado da arte, são apresentadas as abordagens que utilizam grades de ocupação semânticas para a detecção de pessoas e destacadas as diferenças para a abordagem proposta.

3.1 Detecção de pessoas com laser

Na detecção de pessoas com laser, há abordagens que realizam a extração de características geométricas tais como tamanho da borda, convexidade, número de pontos, linhas, quinas, etc. As características são utilizadas para treinar classificadores ou para determinar limiares, como nos trabalhos de Spinello e Siegwart [2008], Premebida et al. [2009], Varvadoukas et al. [2012] e Mozos et al. [2010]. Há também outros trabalhos baseados em casamento de padrões, por exemplo em [Oliveira et al., 2010b], [Pereira et al., 2013] e [Bellotto e Hu, 2009]. Por outro lado, as abordagens que não são baseadas em características utilizam busca por mínimos locais, detecção baseada em movimentos ou subtração do fundo [Cui et al., 2005].

Neste trabalho (Capítulo 5) são utilizados os algoritmos de Bellotto e Hu [2009] e Spinello e Siegwart [2008], por estarem livremente disponíveis na internet e também por apresentarem resultados satisfatórios comprovados em outros trabalhos como [Pereira et al., 2013] e [Varvadoukas et al., 2012]. Ambos os métodos extraem as características necessárias para a detecção a partir de apenas uma leitura do laser, independentemente do movimento dos objetos da cena.

O trabalho de Bellotto e Hu [2009] realiza a detecção de pernas em ambiente interno e é baseado no reconhecimento de padrões típicos de pernas, como pernas separadas, pernas em posição de caminhar e pernas juntas. Ini-

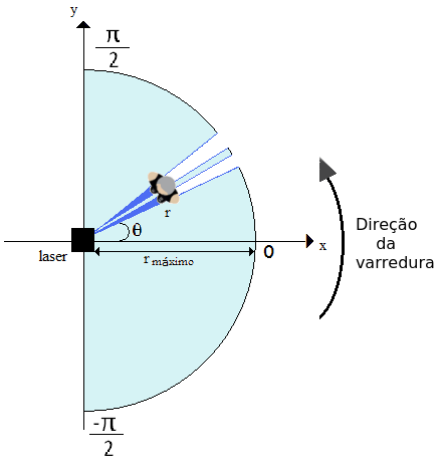


Figura 3.1: Ilustração do funcionamento do laser.

cialmente, são extraídas características de uma leitura do laser e identificados padrões correspondentes às posições das pernas, sendo que o laser encontra-se um pouco acima da altura dos pés. Os dados do laser são processados para encontrar descontinuidades que indicam a presença de objetos na cena. E, por fim, os objetos encontrados são filtrados de acordo com a largura relativa à distância.

Supondo que a resolução angular é constante e que as leituras do laser são armazenadas em um vetor $S = [r_1, ..., r_i, ..., r_M]$, onde r_i é a distância medida na direção θ_i e M é o número total de distâncias lidas (Figura 3.1), a primeira etapa do processamento consiste em aplicar um operador de minimização local, que atribui a cada elemento do vetor o valor mínimo entre todos os elementos de uma vizinhança de determinado tamanho. Como resultado, são removidos possíveis picos causados por reflexão de superfícies inclinadas. Em seguida é aplicado um operador de maximização local, que atribui a cada elemento do vetor o valor máximo entre todos os elementos da vizinhança, para descartar objetos muito finos tais como pés de mesas.

Após o pré-processamento dos dados do laser, o vetor resultante $\hat{S}$ é

usado para detectar descontinuidades (ou bordas verticais, como descrito pelos autores, tendo como referência os pontos do laser interligados e vistos de cima). Considerando uma representação do vetor $\hat{S}$ em um plano cartesiano com ângulos indexados por i no eixo das absissas e as distâncias no eixo das ordenadas, o par $\hat{r}_i, \hat{r}_{i+1}$ pode ser considerado uma borda vertical aproximada se a distância $|\hat{r}_{i+1} - \hat{r}_i|$ for maior que um determinado limiar.

Se $\hat{r}_i > \hat{r}_{i+1}$ considera-se que há uma borda esquerda (L) e se $\hat{r}_i < \hat{r}_{i+1}$ então é uma borda direita (R). Essa classificação das bordas gera uma lista de bordas conectadas e, de acordo com algumas restrições e relações espaciais entre as bordas (distância máxima entre as pernas e limites do tamanho das pernas), os padrões das posições das pernas são buscados conforme determinado a seguir:

- Pernas separadas: L, R, L, R;
- Pernas em posição de caminhar: L, L, R ou L, R, R;
- Pernas juntas: L, R.

Uma ilustração desses padrões é mostrada na Figura 3.2. Os padrões encontrados que atenderem às restrições de dimensão (largura da perna, largura do passo e largura de duas pernas juntas) são considerados pernas de pessoas. Calcula-se então a distância e direção das pernas detectadas no ponto médio de cada padrão. Os resultados apresentaram uma taxa de erro de

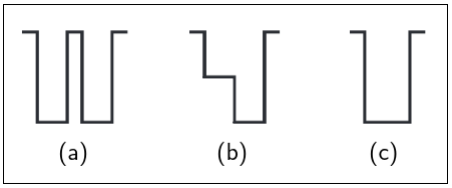


Figura 3.2: Ilustração dos padrões das posições das pernas. (a) Pernas separadas. (b) Pernas em posição de caminhar. (c) Pernas juntas.

aproximadamente 30% em experimentos com o sensor parado e até 3 pessoas simultaneamente em ambiente interno. Uma das desvantagens da abordagem é que os resultados se deterioram com o sensor em movimento, mesmo a baixas velocidades.

Já no trabalho de Spinello e Siegwart [2008] são calculadas características geométricas da leitura do laser que são utilizadas na etapa de treinamento com uma técnica de aprendizado supervisionado. Embora os dados do laser contenham pouca informação sobre pessoas em comparação com a imagem de uma câmera, devido à sua leitura ser tipicamente restrita a um plano de leitura 2D, os pontos associados às pessoas possuem certas propriedades geométricas tais como tamanho da borda, convexidade, número de pontos, etc.

A metodologia de Spinello e Siegwart [2008] está dividida em duas etapas: treinamento e detecção. Na etapa de treinamento é utilizado Adaboost com um conjunto de SVMs lineares para treinamento dos classificadores. Inicialmente, os dados do laser são segmentados e rotulados manualmente como pessoas ou não pessoas. Os segmentos então são descritos por meio de características e o conjunto de características é utilizado pelos classificadores para construir um modelo.

Na etapa de detecção, os dados da leitura do laser são agrupados com base na segmentação proposta pelos autores e avaliados usando o classificador treinado. A técnica de segmentação é baseada na distância Euclidiana entre os pontos: se a distância entre dois pontos adjacentes excede um determinado limiar, então um novo grupo (*cluster*) é gerado. Entretanto, devido ao desempenho insatisfatório desta técnica em ambientes externos e mais complexos, a segmentação proposta também leva em consideração a distância entre os grupos. Após a segmentação, é realizada a descrição geométrica

e estatística dos grupos por meio de características, formando um conjunto de 50 características, dentre as quais incluem-se: o número de pontos, altura, comprimento das bordas, circularidade, diferença angular média, desvio padrão, largura, linearidade, raio, curvatura média, histograma, regularidade das bordas, etc. Em seguida, as características são fornecidas ao modelo obtido na etapa de treinamento para que sejam avaliadas e os grupos são então classificados por meio de um limiar aplicado à probabilidade do grupo estar relacionado a uma pessoa dadas as características analisadas. Essa probabilidade é fornecida pelo classificador. A Figura 3.3 mostra os dados do laser plotados em um gráfico e as pessoas detectadas em destaque.

Experimentos em dois ambientes externos distintos apresentaram taxas de falsos negativos entre 8% e 35%. No ambiente mais complexo, com muitos objetos na cena e múltiplos pedestres em um espaço pequeno, o padrão formado pelos pontos do laser é distorcido, prejudicando a detecção.

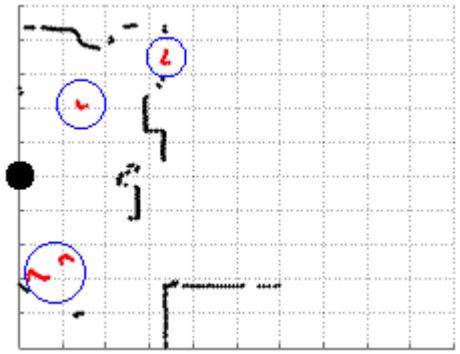


Figura 3.3: Ilustração da detecção de pessoas com laser utilizando o algoritmo de Spinello e Siegwart [2008]. Os dados de um laser 2D estão plotados em um gráfico e as pessoas detectadas são destacadas pelas circunferências. A posição do laser é marcada por um círculo preto à esquerda do gráfico.

3.2 Detecção de pessoas com câmera

A detecção de pessoas em imagens é uma tarefa bastante desafiadora devido à imensa variedade de aparências e poses em que as pessoas podem ser registradas. Na última década surgiram mais de 40 detectores de pessoas em imagens, segundo Benenson et al. [2015], mostrando um notável progresso nos detectores. As abordagens são geralmente baseadas em janelas deslizantes, segmentação ou regiões de interesse [Dollar et al., 2012; Varga et al., 2014].

Há dois trabalhos conhecidos que fundamentam vários dos detectores modernos: detecção de faces utilizando métodos baseados no trabalho de Viola e Jones [2001] e detecção de pessoas utilizando características baseadas em HOG [Dalal e Triggs, 2005]. O método de Viola e Jones [2001] utiliza um conjunto de características simples baseadas nas características de Haar usadas para produzir uma cascata de classificadores fracos treinados por AdaBoost. Esse método é aplicado nos trabalhos de Bellotto e Hu [2009], Pereira et al. [2013] e Araújo et al. [2011].

O trabalho de Dalal e Triggs [2005] realiza o reconhecimento de pessoas e objetos mesmo em ambientes complexos e sob condições de iluminação variáveis utilizando descritores HOG com o classificador linear SVM (*Support Vector Machines*), técnica que passou a ser bastante utilizada na literatura. A ideia básica é que a aparência local e a forma das pessoas podem ser caracterizadas pela distribuição local de gradientes de intensidade ou direção das bordas.

Os detectores de pessoas do estado da arte são baseados em janelas deslizantes sobre pirâmides de características, que são representações da imagem em múltiplas escalas de construção rápida, permitindo execução em tempo real, como por exemplo recentes variações do trabalho de Dollar et al. [2014]. Esse trabalho propõe um algoritmo de detecção de pessoas chamado ACF

(*Aggregated Channel Features*) que é rápido e eficaz, pois consegue realizar o cálculo de pirâmides de características finamente amostradas de forma rápida baseado no fato de que características de múltiplas resoluções da imagem podem ser aproximadas pela extrapolação de escalas próximas, em vez de serem calculadas explicitamente.

Neste algoritmo, a imagem de entrada é transformada em um conjunto de canais de características (ou mapa de características): magnitude normalizada dos gradientes, histograma de gradientes orientados e canais de cor LUV (LUV é um espaço de cores em que duas cores igualmente distantes no espaço de cor de acordo com a métrica Euclidiana são igualmente distantes perceptualmente. Os canais L, U e V correspondem aproximadamente a luminância, verde-vermelho, azul-amarelo.). Os canais são divididos em blocos retangulares e os pixeis de cada bloco são somados. Os canais são então suavizados, formando o vetor de características que e é usado como entrada em uma árvore de decisão construída pelo Adaboost. Os nós da árvore são comparações simples entre o valor da característica e um limiar determinado pelo aprendizado. O Adaboost é usado tanto para seleção de características e também para aprendizado dos limiares nos nós da árvore [Zhang et al., 2015] e uma abordagem de janela deslizante multi-escala é aplicada.

3.3 Fusão sensorial

Existem vários trabalhos de detecção de pedestres que utilizam simultaneamente dados visuais e lasers. De acordo com Oliveira et al. [2010b] as duas abordagens mais comuns para a fusão de laser e câmera são (Figura 3.4):

1. Integração de características dos sensores (fusão a nível de características) ou dos resultados de classificadores (fusão a nível de classificado-

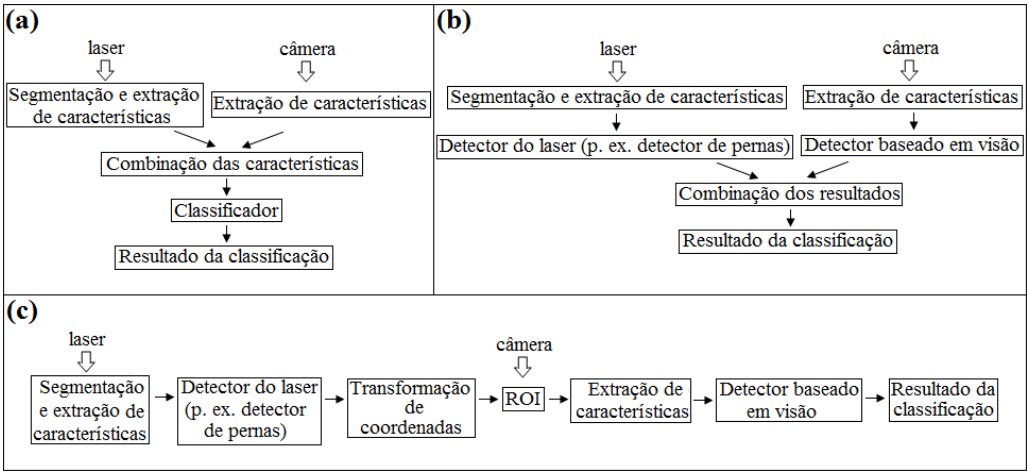


Figura 3.4: Abordagens para fusão sensorial. Abordagem 1: (a) Fusão a nível de características e (b) Fusão a nível de classificadores. Abordagem 2: (c) Algoritmos de visão e do laser executados de forma sequencial.

res), assumindo que a probabilidade de encontrar um objeto é identicamente distribuída e independente em ambos os espaços sensoriais.

2. Uma região de interesse ou ROI (*region of interest*) é determinada por segmentação dos dados do laser e um classificador baseado em visão é utilizado para rotular a ROI projetada.

Dentro da primeira abordagem de fusão de laser e câmera, pode-se citar os trabalhos de Pereira et al. [2013], Araújo et al. [2011], Oliveira et al. [2010b], Gidel et al. [2009], Bellotto e Hu [2009] e Cui et al. [2005]. Nesses trabalhos, as informações complementares e redundantes dos dois sensores são exploradas para maximizar os níveis de confiança e inferência da detecção de pessoas. Na segunda abordagem de fusão, os dados provenientes da leitura do laser são determinantes para que a imagem seja avaliada, ou seja, em situações nas quais o laser falhar na detecção de uma pessoa não haverá detecção na imagem pois não haverá ROI. Pode-se citar os trabalhos de Huerta et al. [2014], Vu et al. [2014], Wu et al. [2011], Broggi et al. [2009], Premebida et al. [2009]

e Spinello e Siegwart [2008] que baseiam-se na segunda abordagem.

Entre os trabalhos que utilizam a primeira abordagem, o trabalho de Cui et al. [2005] realiza a detecção e o rastreamento de pedestres em ambiente interno. Relata-se o rastreamento de mais de dez pedestres simultaneamente em tempo real utilizando dois lasers e uma câmera. A detecção de pedestres é realizada pelo laser, que rastreia a trajetória com base em informações de distância e de forma independente pela câmera, de acordo com histograma de cores. As duas informações de rastreamento são combinadas utilizando o Filtro de Kalman. Inicialmente, objetos em movimento são obtidos dos dados dos lasers pela subtração do fundo, sendo que o modelo do ambiente é conhecido. Os pontos em movimento são integrados em um sistema global de coordenadas e agrupados (*clustering*) em raios menores que 15 cm (largura da perna). Uma vez que um pedestre é detectado pelo laser, a região de seu corpo correspondente na imagem é localizada por meio da calibração entre os sensores e modelada como uma elipse, levando-se em conta o tamanho médio de um ser humano. O modelo de sua aparência é calculado e comparado a outros objetos já rastreados usando o algorimo *mean-shift*, que busca a região mais similar comparada com uma dada região em termos de distribuição de cor.

Em ambientes internos, pode-se citar o trabalho [Bellotto e Hu, 2009], que realiza a fusão de informações de um detector de pernas no espaço do laser com detecção de faces no espaço da câmera. O algoritmo de detecção de pernas foi explicado na Seção 3.1. A detecção de faces é baseada no método de Viola-Jones e a fusão usa o Filtro de Kalman Unscented, que realiza o rastreamento das pessoas.

Nem sempre a fusão de dados de laser e câmera é bem sucedida quando há oclusão parcial de pedestres. Com o objetivo de contornar essa limitação,

no trabalho de Oliveira et al. [2010b] são consideradas informações semânticas obtidas dos dois sensores e a fusão é baseada na relação espacial de classificadores baseados em partes, computada por uma rede lógica de Markov (*Markov logic network*). A segmentação do laser é feita por meio do agrupamento de segmentos, rotulando-os de acordo com um casamento de padrões de partes do corpo humano. Uma janela deslizante no sistema de referência do laser é mapeada na imagem. As janelas projetadas são classificadas por um detector baseado em partes (*parts-based detector*) formado por um conjunto de classificadores denominado HLSM-FINT. Esse conjunto de classificadores utiliza HOG e determinadas características locais (LRF, *local receptive features*) que são classificados por Máquinas de Suporte Vetoriais (SVM) e perceptrons multi-camadas (MLP) e posteriormente combinados em um módulo de fusão baseado em *fuzzy integral* (FINT), que leva em consideração uma medida de inter-relação dos classificadores baseada na importância individual de cada classificador [Oliveira et al., 2010a].

Gidel et al. [2009] combina um laser e duas câmeras em uma abordagem bayesiana para realizar a detecção de pedestres e o rastreamento usando Filtro de Partículas, que são realizados independentemente. A fusão dos dados ocorre pela associação dos pedestres rastreados a partículas 2D e recebem um fator de confiança na fusão de dados que é utilizado na associação de dados.

No trabalho de Araújo et al. [2011], os pedestres são detectados separadamente por cada sensor. Para a segmentação dos dados do laser utilizou-se o método *Point Distance Based* (PDBS), que consiste em transformar pontos próximos entre si em um segmento de reta utilizando um limiar de distância. A detecção e classificação dos dados da câmera utilizou o método de Viola-Jones, que consiste em classificadores de características de Haar otimizados

pelo algoritmo AdaBoost. Os resultados da classificação dos dados de cada sensor foram utilizados para gerar hipóteses de obstáculos, verificadas através da correlação dessas hipóteses. Dessa forma, a fusão consistiu na verificação do centro geométrico do pedestre detectado pelo classificador da câmera em relação à área delimitada pela detecção do laser.

Em [Pereira et al., 2013] o objetivo é cooperação entre uma pessoa e robô e para isso são utilizados uma câmera e sensor a laser, combinando um método de detecção facial e um método de detecção de pernas. Para detecção de faces, foi implementado um método baseado em [Viola e Jones, 2001], em que apenas regiões da imagem onde foi detectada cor da pele são processadas. A detecção de uma pessoa é confirmada caso haja um par de pernas e uma face na mesma região. Este trabalho considera que a pessoa deve estar próxima ao laser (menos de 2 metros) e que a pessoa está voltada para o robô e posicionada na frente do mesmo, admitindo-se que pessoas que queiram interagir com o robô atenderiam essas condições.

Dentro da segunda abordagem de fusão de laser e câmera, pode-se citar o trabalho de Spinello e Siegwart [2008]. A fusão dos dados de laser e câmera consiste em projetar na imagem os *clusters* do laser detectados como pernas de pessoas e então avaliar cada região da imagem correspondente ao *cluster* projetado com um detector baseado em HOG/SVM, com foco no corpo inteiro. Os resultados dos dois detectores são combinados por meio de uma abordagem Bayesiana.

No trabalho de Premebida et al. [2009], os autores apresentam dois métodos de fusão, um centralizado a nível de características (características no espaço do laser e da câmera são combinadas em um único vetor de características e posteriormente classificado por apenas um classificador) e um descentralizado, que utiliza um classificador por espaço de características. No

espaço do laser é realizada uma segmentação e a extração de 15 características distintas baseadas nos pontos do laser. A seguir calcula-se a ROI no espaço da imagem, obtendo as características HOG e covariância. São testados vários classificadores para a fusão, tais como Naive Bayes, GMMC, MCI-NN, FLDA e SVM. Uma abordagem semelhante é apresentada em [Premebida e Nunes, 2013], porém uma estratégia bayesiana é usada para combinar as informações dos sensores com as informações digitais de uma mapa do ambiente, que foi rotulado de forma a destacar regiões em que a presença de pedestres é mais provável (por exemplo, em faixas de pedestres e pontos de ônibus). A segmentação dos dados do laser em conjunto com as informações contextuais geram candidatos que são projetados na imagem. Dentro de cada ROI, um classificador baseado em uma cascata de SVM é usado em forma de detectores baseados em janelas deslizantes em múltiplas escalas.

Um método para detectar pedestres apenas em áreas críticas é apresentado em [Broggi et al., 2009]. A primeira etapa do método é analizar o ambiente usando dados do laser e buscar possíveis pedestres em posições específicas, tais como pedestres prestes a cruzarem a via em frente a veículos estacionados, estando inicialmente parcialmente oclusos. Nessa situação a oclusão parcial do pedestre faz com que o laser não detecte sua presença. Dois classificadores diferentes são usados para agrupar pontos do laser, um para classificar obstáculos e outro para verificar a posição e velocidade dos obstáculos, sendo possível diferenciar obstáculos em movimento de obstáculos estáticos. Obstáculos estáticos (como carros estacionados) são usados para localizar áreas de interesse que serão mapeadas para o referencial da câmera para serem validadas pelo sistema de visão. As janelas mapeadas na imagem são classificadas utilizando AdaBoost com características de Haar (método de Viola-Jones).

Em [Wu et al., 2011], inicialmente pontos do laser são agrupados em segmentos, considerados candidatos a pedestres. Esses segmentos são calculados por um método de agrupamento hierárquico baseado em limiares de distância entre os pontos. Os segmentos são projetados na imagem formando regiões de interesse (ROI), que são classificadas por SVM com base em características de HOG extraídas da ROI. São gerados dois modelos SVM, correspondendo a pedestres com orientação frontal e lateral.

O trabalho de Vu et al. [2014] realiza a fusão de laser, câmera e radar para detectar, rastrear e classificar objetos (pedestres, bicicletas, motos, carros, caminhões). Uma grade de ocupação é usada para identificar objetos estáticos e dinâmicos nos dados do laser e os alvos detectados pelo laser e radar são usados para selecionar regiões de interesse na imagem e, dessa forma, acelerar a detecção de imagens. A fusão combina as informações dos sensores em nível de classificadores usando a teoria de Dempster-Shafer.

No trabalho de Huerta et al. [2014], as detecções de laser e câmera são combinadas em uma abordagem de fusão que considera as relações espaciais e temporais fornecidas por rastreamento utilizando o Filtro de Partículas. No rastreamento, as partículas consistem da posição e velocidade, e a informação temporal obtida realimenta o sistema de visão, o que permite recuperar detecções mesmo quando os sensores falham.

Detalhes adicionais sobre os trabalhos citados podem ser vistos na Tabela 3.1 e na Tabela 3.2.

O laser pode falhar por diversas razões que incluem baixa refletividade de roupas pretas a grandes distâncias [Oliveira et al., 2010b], reflexões e efeitos da luz direta do sol [Spinello e Siegwart, 2008]. Oclusões também podem causar falhas no laser, por exemplo quando uma pessoa está atrás de uma

Trabalho	Câmera	Laser	Sistema de visão	Sistema de laser	Fusão
[Cui et al. 2005]	Lentes de 4,5 mm, ângulo diagonal de 98 graus, slant angle 30-45 graus em direção ao solo	LD-A da IBEO Lasertechnik, 270 graus, 1080 pontos, alcance de 70 m, 30HZ, posicionado no chão (16 cm acima da superfície).	O pedestre detectado pelo laser é localizado na imagem por meio da calibração entre os sensores e modelada como uma elipse. O modelo de sua aparência é calculado e comparado a outros objetos já rastreados usando o algorimo mean-shift.	Detecção de objetos em movimento pela subtração do fundo. Os pontos em movimento são integrados em um sistema global de coordenadas e agrupados em raios menores que 15 cm (largura da perna).	Filtro de Kalman.
[Broggi et al. 2009]	AVT Guppy F-036B, razão de aspecto próxima de 15/9, 15 fps, posicionada dentro da cabine do veículo próxima ao retrovisor, sensibilidade para infra-vermelho próximo, pois foram instaladas lâmpadas emissoras de infra-vermelho na frente do veículo.	SICK LMS211-30206, FOV de 100 graus, resolução angular de 0,25 grau, alcance de 80 m, correção de neblina, interlacing system, 20 Hz.	AdaBoost com características de Haar.	Dois classificadores diferentes são usados para agrupar pontos do laser, um para classificar obstáculos e outro para verificar a posição e velocidade dos obstáculos.	Áreas de interesse localizadas pelo sistema de laser são mapeadas para o referencial da câmera para serem validadas pelo sistema de visão.
[Premebida et al. 2009]	Allied Guppy, campo de visão de 67 graus aproximadamente.	IBEO Alasca-XT, 4 camadas, 12,5 Hz, FOV de 180 graus, resolução angular de 0,5 grau, resolução vertical [-1,2 -0,4 0,4 1.2] graus, alcance de 30 m.	Características HOG e covariância e classificadores FLDA, RBF-SVM e MCI-NN.	Segmentação e a extração de 15 características distintas baseadas nos pontos do laser. Classificadores utilizados foram: Naive Bayes, GMMC, FLDA, RBF-SVM e MCI-NN.	Dois métodos: um centralizado a nível de características (características no espaço dos dois sensores são combinadas em um único vetor e posteriormente classificado por apenas um classificador) e um descentralizado, que utiliza um classificador por espaço de características.
[Bellotto e Hu 2009]	Câmera Pan-Tilt-Zoom (PTZ), posicionada a 1,5 m do solo, 10 Hz (Pioneer) e 25 Hz (Scitos).	Sick LMS211-30206, resolução de ± 1 cm, resolução angular de 0,5 grau, alcance de 80 m, 5 Hz (Pioneer) e 20 Hz (Scitos).	Detecção de faces baseada no método de Viola-Jones	Detecção de pernas busca padrões de pernas juntas (SL), pernas abertas (FS) ou pernas em posição de caminhar (LA).	Filtro de Kalman unscented (UKF), que realiza o rastreamento das pessoas.
[Gidel et al. 2009]	Duas câmeras de vídeo SMAL.	IBEO Alasca-XT.	-	-	Detecção de pedestres ocorre separadamente para cada sensor e também o rastreamento (FP). Os pedestres rastreados são combinados e recebem um fator de confiança na fusão.
[Oliveira et al. 2010]	Campo de visão de 45 graus, 15 fps.	Posicionado na altura da cintura de uma pessoa (0,9 m), alcance de 2 a 20 m, campo de visão (FOV) de 100 graus, resolução angular de 0,25 grau.	Uma janela deslizante no sistema de referência do laser é mapeada na imagem. As janelas projetadas são classificadas por um detector baseado em partes formado por um conjunto de classificadores denominado HLSM-FINT.	A segmentação do laser é feita por meio do agrupamento de segmentos, rotulando-os de acordo com um casamento de padrões de partes do corpo humano.	Rede lógica de Markov.
[Araújo et al. 2011]	-	SICK LMS291, FOV de 100 graus, resolução angular de 1 grau, 181 pontos, alcance de interesse 8 m. Posicionado na altura da cintura.	Método de Viola-Jones, que consiste em classificadores de características de Haar otimizados pelo algoritmo AdaBoost.	A segmentação dos dados do laser utilizou o método Point Distance Based (PDBS), que consiste em transformar pontos próximos entre si em um segmento de reta utilizando um limiar de distância.	Verificação do centro geométrico do obstáculo detectado pelo classificador da câmera em relação à área delimitada pela detecção do laser.
[Wu et al. 2011]	Posicionada no topo do veículo, campo de visão de 45 graus.	SICK LMS291, FOV de 100 graus, resolução angular de 0,25 grau, alcance de 80 m. Posicionado na altura da cintura.	Imagens são classificadas utilizando SVM com base em características de HOG extraídas da ROI. São gerados dois modelos SVM, correspondendo a pedestres com orientação frontal e lateral.	Pontos do laser são agrupados em segmentos, considerados candidatos a pedestres. Esses segmentos são calculados por um método de agrupamento hierárquico baseado em limiares de distância entre os pontos.	Os segmentos do laser são projetados na imagem formando regiões de interesse (ROI), que são classicadas pelo sistema de visão.
[Premebida e Nunes 2013]	Allied Guppy, campo de visão de 67 graus aproximadamente.	IBEO Alasca-XT, 4 camadas, 12,5 Hz, FOV de 180 graus, resolução angular de 0,5 grau, resolução vertical [-1,2 -0,4 0,4 1.2] graus, alcance de 30 m.	Características HOG e cascatas de SVM.	Segmentação.	A segmentação dos dados do laser em conjunto com as informações de um mapa do ambiente geram ROIs na imagem. A fusão é bayesiana.
[Vu et al. 2014]	Câmera monocular.	1 laser.	AdaBoost com características baseadas em HOG.	Uma grade de ocupação é usada para identificar objetos estáticos e dinâmicos nos dados do laser.	Os objetos detectados pelo laser e radar são usados para selecionar ROIs na imagem. A fusão ocorre em nível de classificadores usando a teoria de Dempster-Shafer.
[Huerta et al. 2014]	Câmera PointGrey (resolução de 1024x768 pixesl e 45 graus de FOV).	Laser SICK 2D (FOV = 100 graus).	Características HOG e SVM.	Características geométricas e Adaboost.	Detecções do laser e câmera são combinadas em uma abordagem de fusão que considera as relações espaciais e temporais fornecidas por rastreamento utilizando o Filtro de Partículas.

Tabela 3.1: Detalhes adicionais sobre os trabalhos citados na revisão bibliográfica (primeira parte).

Trabalho	Velocidade do sistema	Tempo de detecção	Resultados	Restrições / observações	Base de dados	Calibração
[Cui et al. 2005]	Parado.	Tempo real.	Deixou de detectar 6 pessoas em 167	Sensores são fixos.	10 minutos de duração, 167 pessoas, ambiente interno.	Modelo de Tsai.
[Broggi et al. 2009]	Máximo de 50 Km/h.	Próximo de 0,001 s por área analisada.	37 detecções corretas, 11 falsos negativos, 1 falso positivo.	Senrores adicionais de taxa de yaw, acelerômetro para computar ângulos de pitch.	Treino do classificador de imagens: 100.000 imagens coletadas e manualmente divididas em classes, sendo usadas também imagens espelhadas com baixa similaridade. Teste: sequências totalizando 10 h, obtidas ao longo de 236 Km em ambientes urbanos, aproximadamente 540.000 imagens. Ambiente externo.	-
[Premebida et al. 2009]	-	-	Melhor acurácia 89,92%.	Sensores fixos em carro elétrico.	ISR-UC_train (1.100 frames com 550 positivos e 550 negativos) e ISR-UC_test (1.400 frames com 400 positivos e 1.000 negativos). Coletados no campus de uma universidade, ambiente externo. Segmentos extraídos do laser foram manualmente validados. O conjunto de imagens foi automaticamente extraído pelo mapeamento das ROI detectadas pelo laser.	Método de Zhang e Pless.
[Bellotto e Hu 2009]	Maior que 0,5 m/s e taxa de rotação de 45 graus/s.	0.25 s.	Falsos positivos: 6,8% (LA), 15% (FS), 70% (SL) e 9% (faces).	Aplicação em robôs de serviço. Os robôs utilizados foram o Pioneer 2 e uma plataforma Scitos G5. Experimentos com até 3 pessoas na cena.	Ambiente interno. Coletada dentro de laboratório, escritório, corredor e arena para robôs. Robô parado: 813 leituras do laser com 1067 pessoas contadas manualmente. Robô em movimento: 619 leituras com 802 pessoas anotadas.	Baseada na posição dos sensores na plataforma robótica.
[Gidel et al. 2009]	De 0 a 50 Km/h.	-	Falsos positivos: 53,6% (laser), 27,4% (câmera), 10,8% (fusão). Verdadeiros positivos: 91,6% (laser), 70,2% (câmera), 92,8% (fusão).	Sensores fixos em veículo da Renault.	Ambientes externos urbanos e estacionamento de veículos.	-
[Oliveira et al. 2010]	Máximo de 30 km/h.	5 s por frame.	Taxa de acerto de 80% a taxa 0,5 falsos positivos por frame (FPPF).	Sensores fixos em carro elétrico.	Treino: INRIA para treinar o detector baseado em partes (2416 pedestres e 15000 não-pedestres. Base coletada no campus da universidade com 2672 frames para treino e 2157 para teste. Imagens manualmente anotadas e detecção do laser avaliada no espaço da imagem. Dados coletados à velocidade máxima de 9,7 Km/h. Ambiente externo.	Método de Zhang e Pless.
[Araújo et al. 2011]	-	-	Acurácia de 72,5%, taxa de verdadeiros positivos de 49,2% e taxa de falsos positivos de 0,18%.	-	Classificador de imagem foi treinado com base do INRIA.Foram utilizadas 1023 imagens com pedestres e suas versões espelhadas e 4140 imagens sem pedestres.	Relações geométricas entre um ponto detectado pelo laser que está dentro do campo de visão da câmera.
[Wu et al. 2011]	Máximo de 30 km/h.	0,158 s por frame.	Taxa de acerto próxima de 70% a taxa 0,45 falsos positivos por frame.	-	Treino: foram coletadas mais de 5000 pedestres e 5000 não pedestres. Teste: ISR-UC-imglidar-svnc com 2157 frames e leituras do laser sincronizadas.	Método de Zhang e Pless.
[Premebida e Nunes 2013]	-	-	Curvas Taxa de Deteção x FPPF (por exemplo, para FPPF = 4, a taxa de detecção é cerca de 76%).	Sensores fixos em carro elétrico (ISRobotCar).	LIPD	Métodos de Bouguet, Vasconcelos e Heikkila e Silven.
[Vu et al. 2014]	-	Tempo médio de 40 ms.	Taxa de detecção de pedestres entre 82% e 97% e taxa de falsos positivos igual a 12%.	Sensores fixos em veículo (CRF).	Ambiente externo urbano.	-
[Huerta et al. 2014]	Velocidade máxima de 30 Km/h.	7 fps.	Taxa de acerto de 84,13% a 0,5 FPPF.	Sensores fixos em carro elétrico.	ISR-UC-imglidar-sync	Método de Zhang e Pless.

Tabela 3.2: Detalhes adicionais sobre os trabalhos citados na revisão bibliográfica (segunda parte).

caixa de tamanho médio, a câmera consegue visualizar a parte superior do seu corpo, mas um laser posicionado na altura das pernas somente detectará a caixa. Dessa forma, as técnicas de fusão baseadas na segunda abordagem (em que uma ROI é determinada por segmentação dos dados do laser e posteriormente é classificada utilizando sua projeção na imagem) tendem a deixar de detectar um maior número de pessoas quando avaliadas em relação à quantidade de pessoas realmente existentes na cena e não apenas aquelas computadas por meio do laser. Em experimentos apresentados nesta tese (Capítulo 5), o comportamento da fusão de uma câmera e de um sensor a laser para a detecção de pessoas foi analisado e observou-se que a câmera ou o laser individualmente estão sujeitos a oclusões que ocorrem de maneira diferente para cada sensor. Dessa forma, a oclusão de uma pessoa para a câmera frequentemente não ocorre no mesmo instante para o laser e vice-versa. Assim, em aplicações como contagem de pessoas, segurança e busca, por exemplo, a utilização de mais de um sensor é importante para a robustez do sistema. Além disso, falhas nos sensores devido a outros fatores fazem a fusão também necessária em outras aplicações tais como navegação de robôs.

Nos trabalhos revisados notam-se determinadas restrições práticas. Por exemplo, geralmente assume-se que o laser está sempre paralelo ao solo - essa restrição será violada quando o sensor mover-se em solos inclinados. Os autores de [Oliveira et al., 2010b] propõem como solução para esse problema a utilização de um laser com múltiplas camadas. Erros de imprecisão da calibração dos sensores também prejudicam a fusão sensorial [Premebida et al., 2009]. Outra restrição observada é que o funcionamento dos sistemas está condicionado a baixas velocidades do sensor. Além disso, a maior parte dos trabalhos podem detectar e rastrear apenas uma ou poucas pessoas [Cui et al., 2005].

Embora sistemas de detecção de pedestres rápidos e eficientes sejam ainda um desafio em aberto [Bengler et al., 2014], observa-se que a fusão sensorial traz vantagens como maior acurácia em relação aos sensores individuais. As abordagens de fusão de alto nível (objetos ou classificadores) têm a vantagem de que parte do processamento é realizado a nível dos sensores, oferecendo modularidade e exigindo menos recursos de comunicação. Adicionalmente, com a fusão das informações também é possível detectar pessoas com um campo de visão mais amplo formado pela união dos campos de visão de todos os sensores do sistema. Além da fusão sensorial, outra forte tendência nos trabalhos atuais é a utilização de informações temporais, possibilitando recuperar detecções mesmo quando os sensores falham. As abordagens bayesianas tratadas na próxima seção em geral incorporam informações temporais, principalmente as abordagens baseadas em variações do Filtro de Bayes.

3.4 Abordagens bayesianas

Há vários trabalhos na literatura que usam abordagens Bayesianas para detecção de pedestres. Entretanto existem diferenças entre essas abordagens, que variam desde a utilização simples da fórmula de Bayes até o uso de técnicas derivadas do Filtro de Bayes (Filtro de Kalman, Filtro de Partículas, Grades de ocupação, etc).

A fórmula de Bayes é em geral utilizada para, dadas as características extraídas e a confiança relativa a elas, calcular a probabilidade de um obstáculo ser pessoa (em problemas de classificação ou segmentação, modelos de sensores) ou a probabilidade de uma pessoa estar em uma determinada posição (configuração), podendo ou não utilizar informações de instantes anteriores incorporadas no termo da probabilidade *a priori* para calcular a probabili-

dade atual. Exemplos de trabalhos que utilizam esta abordagem são: [Kooij et al., 2014; Utasi e Benedek, 2013; Elguebaly e Bouguila, 2011; Bota e Nedesvchi, 2008; Ngako Pangop et al., 2008, 2007; Monteiro et al., 2006; Zhao e Nevatia, 2003].

Por outro lado, as abordagens baseadas em variações do Filtro de Bayes incluem a etapa de predição, baseada em modelos matemáticos e informações sobre o passado, e uma etapa de correção que utiliza dados de sensores para atualizar a predição. Essas abordagens geralmente realizam o rastreamento da posição de características e objetos, fornecendo informação temporal que pode realimentar a etapa de predição ou também etapas de detecção de objetos e fusão de dados. Os trabalhos que seguem essa ideia incluem: [Kim et al., 2015; Gepperth et al., 2014; Huerta et al., 2014; Bota e Nedesvchi, 2008; Cho et al., 2014; Schneider e Gavrila, 2013; Yoder et al., 2014; Gidel et al., 2009; Ngako Pangop et al., 2008, 2007; Linzmeier, 2004].

A metodologia proposta neste trabalho de doutorado pode ser classificada como uma abordagem bayesiana, pertencendo à subclasse de trabalhos que utilizam variações do Filtro de Bayes. Nessa metodologia, o Filtro de Bayes é utilizado para combinar detectores de pessoas, ou seja, a fusão de informações ocorre a nível de classificadores pois as informações combinadas são dados previamente processados e classificados como pessoas por um conjunto de detectores publicados anteriormente na literatura. As detecções resultantes da combinação têm a sua confiança maximizada em relação a detectores de pessoas individualmente e serão fornecidas para as aplicações (por exemplo rastreamento de pessoas, contagem de pessoas, navegação, interação de robôs com humanos, detecção de eventos) por meio de uma grade de ocupação semântica, que inclui informações como posição das pessoas e a probabilidade de existir pessoas nas células da grade, que correspondem a regiões do espaço.

Os trabalhos relacionados a grades de ocupação semântica são resumidos na próxima seção.

3.5 Grades de ocupação semânticas

As grades de ocupação semânticas são usadas em diversas aplicações pois permitem uma compreensão de alto nível do ambiente. A seguir são apresentados alguns trabalhos e suas aplicações.

No trabalho de Wolf e Sukhatme [2008] são desenvolvidas técnicas para construir mapas baseados em grades para representar a atividade e navegabilidade do ambiente utilizando um robô equipado com lasers 2D. A atividade consiste na ocupação do espaço por entidades dinâmicas ao longo do tempo, detectadas por meio de seu movimento. Com base na atividade, as células da grade são classificadas em ruas ou calçadas.

Shi et al. [2010] apresenta uma abordagem para classificar lugares de um ambiente com base em um sensor a laser. A classificação semântica de lugares utiliza uma abordagem probabilística para construção de grades cujas células podem ser rotuladas como corredores, escritórios ou salas de aula.

Bouzouraa e Hofmann [2010] propõe a combinação de mapas de ocupação com rastreamento de objetos por meio da associação de objetos em movimento com células dinâmicas. O mapa e o rastreamento de objetos compartilham uma interface composta por uma lista de objetos rastreados e suas células correspondentes na grade de ocupação. O trabalho realiza a fusão de laser e radar. O laser auxilia na detecção de objetos em movimento comparando-se a sua leitura do ambiente com a grade de ocupação construída com a leitura anterior. As informações de velocidade obtidas com o radar são associadas com a leitura do laser. Veículos são representados na grade,

sendo que a sua detecção é baseada em medidas de referência do radar.

Um mapa semântico de ambiente interno é apresentado por Jebari et al. [2011], contendo informações provenientes de dois lasers, um anel de sonares e uma câmera. O objetivo é detectar objetos como computadores, caixas, jornais, livros, telefones, armas, robôs, etc. A detecção de objetos é realizada, primeiramente, pela segmentação das imagens e em seguida a extração de características e classificação baseada na abordagem de palavras visuais (*bag of visual words*).

Wang e Chen [2011] propõe uma representação de mapa semântico baseada em objetos que utiliza uma linguagem de representação de conhecimento chamada *Web Ontology Language*. O modelo do ambiente é composto por objetos como mesa, parede, cômodo, etc, que são detectados usando visão estéreo. Uma extensão deste trabalho é apresentada em [Filliat et al., 2012], com a adição de um sensor RGBD (Microsoft Kinect) para construir um mapa semântico que inclua informações como cômodos e sua conectividade, os objetos contidos nele e o material das paredes e do chão. A detecção de objetos utiliza uma RNA. O mapa de ocupação é produzido utilizando um laser 2D e um sonar (para detecção de objetos pequenos, vidros e espelhos) com a inclusão dos obstáculos detectados pelo Kinect.

Em [Liu et al., 2012] é proposto um método probabilístico para analisar um modelo semântico do ambiente baseado em uma grade de ocupação. O modelo probabilístico do mundo tem a forma de um grafo e consiste em cômodos e vãos de portas. A grade é gerada por uma técnica de SLAM tradicional.

O trabalho de Liu e von Wichert [2014] propõe um método para extrair uma planta baixa com conceitos abstratos de mapas de ocupação utilizando uma formulação Bayesiana. São usados conceitos como quarto, corredor,

parede, porta, etc. Utiliza-se um laser e considera-se que já existe um mapa construído.

As grades de ocupação semântica que envolvem detecção de pessoas foram usadas anteriormente por Vu et al. [2014]; Hofmann et al. [2011]; Lu et al. [2008]; Linzmeier et al. [2004].

O trabalho de Vu et al. [2014], já citado na Seção 3.3, é uma recente aplicação das grades de ocupação semânticas. Uma restrição deste trabalho é que a fusão sensorial é realizada apenas na interseção dos campos de visão dos sensores, de forma que a classificação das pessoas ocorre somente nas ROIs das imagens que foram determinadas pelas detecções nos dados do radar e do laser.

Em Hofmann et al. [2011], é apresentada uma abordagem para rastreamento e análise de comportamento em ambientes domésticos utilizando vários sensores como câmera, sensores de temperatura, etc. As informações dos sensores são combinadas em uma grade de ocupação 3D. A abordagem apresentada realiza a reconstrução de formas tridimensionais dos objetos, que são obtidas por subtração do fundo e comparadas a padrões simples representando pessoas em pé e na iminência de cair no chão para detecção de eventos. Como essa abordagem baseia-se na subtração do fundo para detectar regiões de interesse, ela não é adequada para ambientes desconhecidos e que possam ocorrer alterações estruturais, além de apresentar maior demanda por recursos computacionais em ambientes de grande dimensão. O movimento do objeto pode ser um indicativo de que existe uma pessoa ali, porém outros objetos também podem mover-se, como animais e veículos. O tamanho do obstáculo e a temperatura do objeto também são considerados em algumas abordagens, mas da mesma forma que o movimento, pode-se confundir pessoas com outras entidades de medidas semelhantes.

Lu et al. [2008] descrevem uma abordagem de mapa de fusão de ocupação (OFM - *Occupancy Fusion Map*) com a fusão de laser e visão estéreo para detecção de pedestres. É proposto um método para construir uma grade de ocupação com visão estéreo como uma alternativa ao problema da dificuldade de detecção de objetos devido a descontinuidades nos dados da grade de ocupação tradicional. A leitura do laser é segmentada pelo agrupamento dos pontos levando-se em consideração sua proximidade, sendo ajustada uma reta ao grupo. Segmentos de reta que não correspondem a determinado tamanho são rejeitados, restando apenas aqueles que podem corresponder a pedestres ou objetos de dimensão similar. Para a seleção de ROIs nas imagens da câmera, uma grade de ocupação é construída com base nas coordenadas X-disparidade em vez de X-Z. Para a detecção de pedestres por partes, *blobs* são detectados e sua largura e altura são analisados. A construção do OFM é realizada pela fusão das ROIs detectadas, inicialmente transformando a localização dos dados do sistema de coordenadas do laser para o sistema X-disparidade e buscando correspondências entre as ROIs detectadas pelo laser e pela câmera. As ROIs são então classificadas como pedestres ou não-pedestres.

O trabalho de Linzmeier et al. [2004] apresenta um sistema de detecção baseado em sensores de infravermelho fixos em um veículo. Como os sensores não fornecem informação de distância, são aplicadas técnicas probabilísticas para calcular a posição dos objetos de acordo com as posições combinadas de sensores cujos campos de visão tem uma interseção. As medidas obtidas são mapeadas para uma grade de ocupação. Os sinais de saída dos sensores são interpretados comparando-se com um sinal de referência. Embora os sensores de infravermelho não interfiram no ambiente, os autores relatam que sua desvantagem é a interpretação não-trivial do sinal quando o veículo

está em movimento e há mudanças no fundo.

Nos trabalhos encontrados na literatura sobre detecção de pessoas utilizando grades semânticas, a detecção de pessoas é baseada em movimento, medidas geométricas e na subtração do fundo. Essas abordagens estão sujeitas a erros devido, por exemplo, à grande variação de medidas das pessoas, desde uma criança com medidas pequenas até um adulto com medidas maiores. Além disso, ambientes reais podem ser complexos e é possível, em várias situações, a presença de obstáculos similares a pessoas. Nesta tese, a implementação da metodologia proposta, explicada no próximo capítulo, contará com uma grade semântica cujas informações são obtidas de uma combinação de classificadores baseados em características específicas de pessoas para detectá-las em ambientes desconhecidos, de modo que a detecção de pessoas seja mais robusta e independente do ambiente em que o sistema está inserido.

Capítulo 4

Metodologia proposta

Este capítulo apresenta a metodologia proposta. Essa metodologia combina múltiplos detectores de pessoas para obter detecções com maior precisão e acurácia do que os detectores separadamente. A metodologia proposta pode ser classificada como uma abordagem bayesiana, dentro do subconjunto de trabalhos que utilizam variações do Filtro de Bayes (Seção 3.4). Com a fusão dos detectores, obtém-se as probabilidades de regiões no espaço estarem ocupadas por pessoas, que podem ser disponibilizadas para aplicações (rastreamento de pessoas, contagem de pessoas, navegação, interação de robôs com humanos, detecção de eventos, etc) por meio de, por exemplo, uma grade de ocupação semântica (Seção 2.3). A implementação da grade de ocupação semântica realizada neste trabalho inclui informações como posição das pessoas e a probabilidade de existir pessoas nas células da grade, que correspondem a regiões do espaço. Um diagrama de blocos da metodologia proposta é mostrado na Figura 4.1.

A metodologia apresentada aplica o Filtro de Bayes (Seção 2.2) para (i) prever a presença de pessoas em uma região específica do espaço e (ii) corrigir essa previsão por meio de informações fornecidas por um conjunto de detectores de pessoas. Neste trabalho a predição é realizada em duas fases, uma baseada no movimento das pessoas e outra baseada no movimento dos

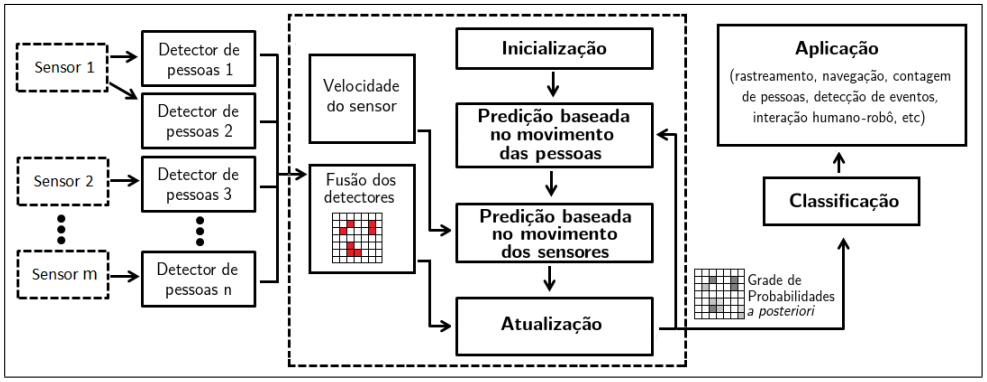


Figura 4.1: Visão geral das etapas da abordagem proposta, destacada pelo maior retângulo tracejado. Adicionalmente, o esquema mostra as outras etapas necessárias para integrar a abordagem a alguma aplicação.

sensores. As seções seguintes detalham cada etapa da metodologia proposta.

4.1 Predição baseada no movimento das pessoas

Seja X a variável aleatória que representa o estado da região centrada nas coordenadas (a, b) em relação ao tipo de objeto que possa estar presente nessa região. Se o espaço em consideração for, por exemplo, uma região de uma rua em frente a um veículo, o experimento de observar esta região no tempo t pode resultar na conclusão de que existe nesta região uma pessoa, ou um carro, ou uma bicicleta, ou um animal, ou outros obstáculos ou a região pode estar livre. Neste caso, a variável aleatória X pode assumir os seguintes valores: pessoa, carro, bicicleta, animal, outros obstáculos ou espaço livre. Portanto, o espaço amostral S consiste de todos os possíveis resultados obtidos da observação tal que $S = \{\text{pessoa, carro, bicicleta, animal, } \ldots, \text{ espaço livre}\}$. De acordo com o segundo axioma da probabilidade [Papoulis e Pillai, 2002], a soma das probabilidades de todos os eventos do espaço amostral é igual a

1, ou seja:

$$\sum_{x \in S} P(X = x) = P(X = \text{pessoa}) + \cdots + P(X = \text{espaço livre}) = 1. \quad (4.1)$$

Como o foco deste trabalho é a detecção de pessoas, considera-se que o espaço amostral está particionado em dois subconjuntos: $S = \text{p} \cup \text{np}$, onde p é o evento $\{X = pessoa\}$ e np (não pessoa) é o evento em que podem ocorrer outros elementos de S, exceto pessoas.

Neste trabalho, para simplificar a notação, a probabilidade do evento p ocorrer ou $P(X = \text{p})$ será denotado por $P(\text{p})$. O cálculo de $P(\text{p})$ depende de diversos fatores que podem ser considerados, tais como: o estado da região no instante da última observação, o deslocamento de pessoas de outras regiões para a região de interesse, a informação dos sensores que observam a região e o movimento desses sensores.

A predição baseada no movimento das pessoas é a etapa da metodologia que calcula a probabilidade de uma região específica do espaço estar ocupada por pessoas com base no conhecimento do movimento das pessoas (e sua velocidade v^{pessoas}) e da probabilidade de existência de pessoas na região e sua vizinhança no instante anterior $(t-1)$. Dessa forma, dada a crença $bel(x_{t-1})$ da presença de pessoas em uma região centrada em (a, b) no tempo anterior $(t-1)$ e o modelo de movimento das pessoas, representado por uma função densidade de probabilidade (PDF), a probabilidade de que a região esteja ocupada no tempo t é calculada pela Equação (4.2). Essa equação é baseada na equação de predição do Filtro de Bayes e será usada a notação $\overline{bel}'(x_t)$ e $\overline{bel}(x_t)$ para diferenciar as duas etapas da predição na metodologia proposta, onde $\overline{bel}'(x_t)$ pode ser escrita como $P(x_t|z_{t-1}, v_t^{\text{pessoas}})$ e corresponde à etapa de predição baseada no movimento das pessoas e $\overline{bel}(x_t)$ pode ser escrita como $P(\text{p}|v_t^{\text{sensores}})\overline{bel}'(x_t = \text{p})$ e corresponde à etapa de predição baseada no

movimento dos sensores combinada com a velocidade das pessoas no tempo atual e a probabilidade *a posteriori* do estado anterior, antes de incorporar a medição dos detectores de pessoas no tempo t.

Logo, $\overline{bel}'(x_t)$ é dada por:

$$\overline{bel}'(x_t) = \int P(x_t|x_{t-1}, v_t^{\text{pessoas}}) bel(x_{t-1})\ dx_{t-1}\,. \tag{4.2}$$

Como o estado x pode assumir apenas dois valores, $x = \text{p}$ ou $x = \text{np}$, o espaço de estados é discreto e a integral em (4.2) torna-se uma soma finita:

$$\overline{bel}'(x_t) = \sum_{x_{t-1}} P(x_t|x_{t-1}, v_t^{\text{pessoas}}) bel(x_{t-1}). \tag{4.3}$$

No tempo t, a região centrada em (a, b) será ocupada por pessoa em duas situações: se existia alguma pessoa em (a, b) no tempo $t-1$ e ela manteve-se parada ou se não existia nenhuma pessoa na região no tempo $t-1$ mas alguma pessoa vinda de outra região (i, j) moveu-se para (a, b). Portanto, a Equação (4.3) torna-se similar à equação usada para calcular a probabilidade de ocupação de regiões por espécies no campo da Ecologia, que também leva em conta o movimento de indivíduos [MacKenzie et al., 2003]. Esse modelo considera a ocupação de áreas (colonização) e sua desocupação (extinção) de forma semelhante a pessoas que movem-se de uma região para outra, porém com diferentes velocidades e probabilidades de colonização e extinção. Com base nesse modelo, a Equação (4.3) pode ser reescrita como:

$$\begin{aligned}\overline{bel}'(x_t = \text{p}) =& P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{pessoas}} = 0) bel(x_{t-1} = \text{p}) + \\ &+ P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{pessoas}} = v^{\text{média}}) bel(x_{t-1} = \text{np})\,,\end{aligned} \tag{4.4}$$

onde:

- $\overline{bel}'(x_t) = P(x_t|z_{t-1}, v_t^{\text{pessoas}})$ é a probabilidade do estado x_t no tempo t

condicionado à medição passada z_{t-1} e a velocidade das pessoas v_t^{pessoas}.

- $bel(x_{t-1}) = P(x_{t-1}|z_{t-1}, v_{t-1}^{\text{pessoas}})$ representa a crença *a priori* sobre o estado x_{t-1}, isto é, a probabilidade do estado x_{t-1} condicionado à medição passada dos sensores z_{t-1} e à velocidade passada v_{t-1}^{pessoas}.

- $P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{pessoas}} = 0)$ é a probabilidade das pessoas não se moverem da sua região atual (a, b).

- $P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{pessoas}} = v^{\text{média}})$ é a probabilidade de pessoas ocuparem a região (a, b) dado que sua velocidade é $v^{\text{média}}$.

A Equação (4.4) calcula a probabilidade de existir pessoa na região centrada em (a, b) pela soma das probabilidades relacionadas ao movimento das pessoas para esta região e à informação de alguma pessoa parada ali no instante anterior. Para outras regiões do espaço, a probabilidade de existir pessoa deve ser calculada separadamente utilizando essa equação para cada região.

A probabilidade $P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{pessoas}} = 0)$ é dada por:

$$P(\text{p}|x_{t-1} = \text{p}, v_t^{\text{pessoas}} = 0) = p_{\text{imóvel}}\,, \tag{4.5}$$

onde $p_{\text{imóvel}}$ é a probabilidade das pessoas permanecerem paradas em suas posições atuais. Por outro lado, $P(\text{p}|x_{t-1} = \text{np}, v_t^{\text{pessoas}} = v^{\text{média}})$ é a probabilidade de existir pessoa em uma região levando-se em consideração a soma das contribuições das regiões vizinhas devido à possibilidade das pessoas se movimentarem com uma determinada velocidade. Essa probabilidade é calculada por:

$$P(\mathrm{p}|x_{t-1}=\mathrm{np}, v_t^{\text{pessoas}} = v^{\text{média}}) = \\ =1 - \frac{p_{\text{livre}}}{p_{\text{livre}} + \sum_{k=(i,j)\in G} \frac{p_{\text{livre}} p_k bel(x_{t-1}^{i,j}=\mathrm{p})}{(1-p_k) bel(x_{t-1}^{i,j}=\mathrm{p})}}\,, \tag{4.6}$$

onde $x_{t-1}^{i,j}$ refere-se ao estado relacionado a uma determinada região do espaço centrada em (i, j) estar ocupada por pessoa ou não pessoa no instante de tempo t, p_{livre} é a probabilidade da região (a, b) permanecer livre de pessoas e p_k é a probabilidade de pessoas moverem-se de (i, j) para (a, b), que pode ser calculada utilizando-se o modelo de movimento das pessoas. Essas probabilidades estão relacionadas por:

$$p_{\text{livre}} = \prod_{k=(i,j)\in G} (1 - p_k)\,. \tag{4.7}$$

O conjunto G nas equações (4.6) e (4.7) é o espaço de todas as regiões que podem ser ocupadas por pessoas. Neste trabalho considera-se que o espaço é discreto. Além disso, em implementações práticas, G pode ser restrito à vizinhança determinada por todas as regiões que uma pessoa pode alcançar a partir de (a, b) em um único intervalo de tempo. Dado que há N regiões do ambiente ocupadas por pessoas, o evento da pessoa q ($1 \leq q \leq N$) ocupar a região (a, b) é mutualmente excludente do evento de uma outra pessoa r ($1 \leq r \leq N$ e $r \neq q$) ocupar a região (a, b), além de serem considerados independentes. Dessa forma, a probabilidade da região (a, b) estar ocupada por alguma das pessoas do ambiente deve considerar $N + 1$ possibilidades, ou seja, a possibilidade de cada uma das N pessoas de outras regiões mover-se para (a, b) e a possibilidade de nenhuma das pessoas mover-se para (a, b), podendo ser calculada pela soma das probabilidades dos eventos citados. Essa probabilidade é obtida primeiramente calculando-se a probabilidade de (a, b) não ser ocupada (p_{livre}), que é normalizada e então subtraída de 1 para obter

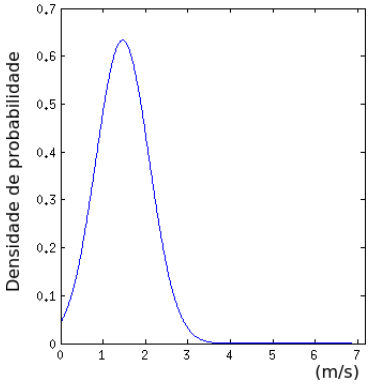


Figura 4.2: Função densidade de probabilidade (PDF) utilizada no modelo de movimento das pessoas.

a probabilidade de (a, b) estar ocupada por uma pessoa. A normalização leva em consideração todas as possibilidades de ocupação da região (a, b), somando suas probabilidades para formar o denominador da Equação (4.6).

O modelo de movimento das pessoas, p_k, descreve a distribuição de velocidades nas quais uma pessoa geralmente caminha. Nesse modelo, o espaço amostral é composto pelas velocidades em que uma pessoa move-se e também inclui a probabilidade de uma pessoa permanecer parada. Há um limite físico para a velocidade máxima alcançável por uma pessoa, que pode ser considerada como a velocidade média do atleta recordista mundial da modalidade atletismo [Hogenboom, 2013], pouco provável para pessoas comuns. Portanto, o modelo do movimento das pessoas deve considerar uma probabilidade baixa para altas velocidades e uma maior probabilidade em torno da velocidade média com a qual elas caminham. A literatura sugere que o modelo de movimento das pessoas pode ser representado por uma PDF normal com média igual a $1,46$ m/s e desvio padrão de $0,63$ [Daamen e Hoogendoorn, 2007]. A Figura 4.2 mostra a curva dessa PDF, com a velocidade da pessoa no eixo x e a densidade de probabilidade no eixo y.

Neste modelo, as pessoas têm maior probabilidade de caminhar com a

velocidade média, e essa probabilidade vai reduzindo conforme a velocidade aumenta demasiadamente ou se aproxima de zero. De forma a considerar também a probabilidade das pessoas estarem paradas, foi proposta uma função que utiliza o modelo gaussiano do movimento das pessoas juntamente com a possibilidade de parada. Considerando que a probabilidade das pessoas estarem paradas em um espaço urbano é $p_{\text{imóvel}}$ (na implementação da metodologia utilizou-se o valor $0,85$, como será explicado no Capítulo 5), o seguinte modelo é proposto:

$$p_k = \begin{cases} p_{\text{imóvel}}, \text{ se } dist((a,b),(i,j)) = 0; \\ (1 - p_{\text{imóvel}}) \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(dist((a,b),(i,j))/\Delta t - \mu)^2}{2\times\sigma^2}}, \text{ se} \\ dist((a,b),(i,j)) \neq 0. \end{cases}$$

onde μ é a média da distribuição, σ é o desvio padrão, Δt é o intervalo de tempo entre t e $t-1$ e a função $dist(\cdot,\cdot)$ calcula a distância entre dois pontos em um espaço bidimensional.

Neste modelo, assume-se que as pessoas podem caminhar livremente em qualquer direção (não há conhecimento sobre obstáculos ou sobre o mapa do ambiente) e portanto considera-se que pessoas podem mover-se em qualquer direção com a mesma probabilidade.

Embora existam diversos modelos de movimento de pessoas na literatura, estes são voltados para aplicações de rastreamento e simulação de multidão. Há modelos simples que consideram a velocidade das pessoas constante, usados em algoritmos de rastreamento que, dadas a posição e a velocidade das pessoas no instante passado, estimam a posição e velocidade atuais e para isso, dependem da associação das pessoas no tempo [Keller e Gavrila, 2014; Schneider e Gavrila, 2013; Ngako Pangop et al., 2007]. Modelos mais sofisticados também são baseados no movimento individual das pessoas, porém

consideram vários termos tais como localização de destinos, colisões com outros agentes, variações na velocidade, efeito de comportamento em grupo, mapas locais, etc. Estes modelos podem depender de treinamento na mesma cena em que o sistema será usado e da utilização de características específicas da cena para determinar parâmetros (como a velocidade das pessoas) [Kim et al., 2015; Antonini et al., 2006]. Nesta tese de doutorado buscou-se a elaboração de um modelo simples e genérico de forma a não ser necessária a associação das pessoas na cena e que não levasse em consideração mapas do local, permitindo seu uso em ambientes desconhecidos.

A Figura 4.3 mostra um exemplo da etapa de predição aplicada repetidamente após um sensor hipotético detectar pessoas no ambiente, que é representado por uma grade semântica com 30×30 células. Inicialmente, o ambiente é desconhecido. A primeira grade (Figura 4.3(a)) mostra uma distribuição uniforme em todas as células, que assumem o valor de probabilidade igual a $0,5$ indicando que não há conhecimento sobre a presença de pessoas. No próximo instante, um sensor hipotético observa a presença de pessoas na região central do ambiente representado pela grade (Figura 4.3(b)), atualizando a grade anterior com essa observação por meio de um mapeamento das coordenadas das pessoas para as células correspondentes da grade, de forma a obter a probabilidade *a posteriori* para cada célula. Considera-se, neste exemplo, que o sensor não possui incertezas associadas à localização das detecções. Quanto mais escura a célula, maior é a probabilidade de existir pessoa. Mesmo nas células em que não houve detecção de pessoas, a probabilidade, embora seja pequena, não é igual a zero devido à incerteza inerente ao processo de sensoriamento (por exemplo, falhas na detecção e falsas detecções) e da mesma forma as células mais escuras não possuem probabilidade igual a 1 nas regiões onde o sensor detectou pessoas. Nos próximos instantes

de tempo, as leituras do sensor não serão consideradas para que se possa observar o efeito da predição nas probabilidades. A partir da Figura 4.3(c), em cada instante de tempo é aplicada a Equação (4.3) considerando como crença *a priori* as probabilidades do instante anterior. A predição baseada no movimento das pessoas tem o efeito de suavizar a crença realizando um espalhamento que reflete a incerteza da situação do ambiente quando não há observação de nenhum sensor, pois as pessoas observadas podem ter se deslocado para regiões vizinhas e a probabilidade de haver pessoas nas células dessas regiões é aumentada. Quanto mais o tempo passa, maior é a incerteza da localização dessas pessoas de forma que as probabilidades das células tendem a estabilizar em um valor uniforme. Observa-se que há um efeito nas bordas da grade que torna essas células levemente mais escuras que as outras células, provocado pela incerteza no espaço fora da grade, em que considera-se a probabilidade de existir pessoas igual a $0,5$.

A próxima subseção descreve o segundo passo da predição da metodologia proposta.

4.2 Predição baseada no movimento dos sensores

No segundo passo da predição é calculada a probabilidade de existência de pessoas em determinada região do espaço em relação ao sistema de coordenadas dos sensores. Quando há movimento dos sensores, as coordenadas de cada região do espaço são deslocadas em relação ao sistema de coordenadas dos sensores. Dessa forma, é necessário modelar o movimento dos sensores para estimar as novas posições das regiões. Assumindo que os sensores são fixos em um robô móvel, o modelo de movimento pode utilizar informações

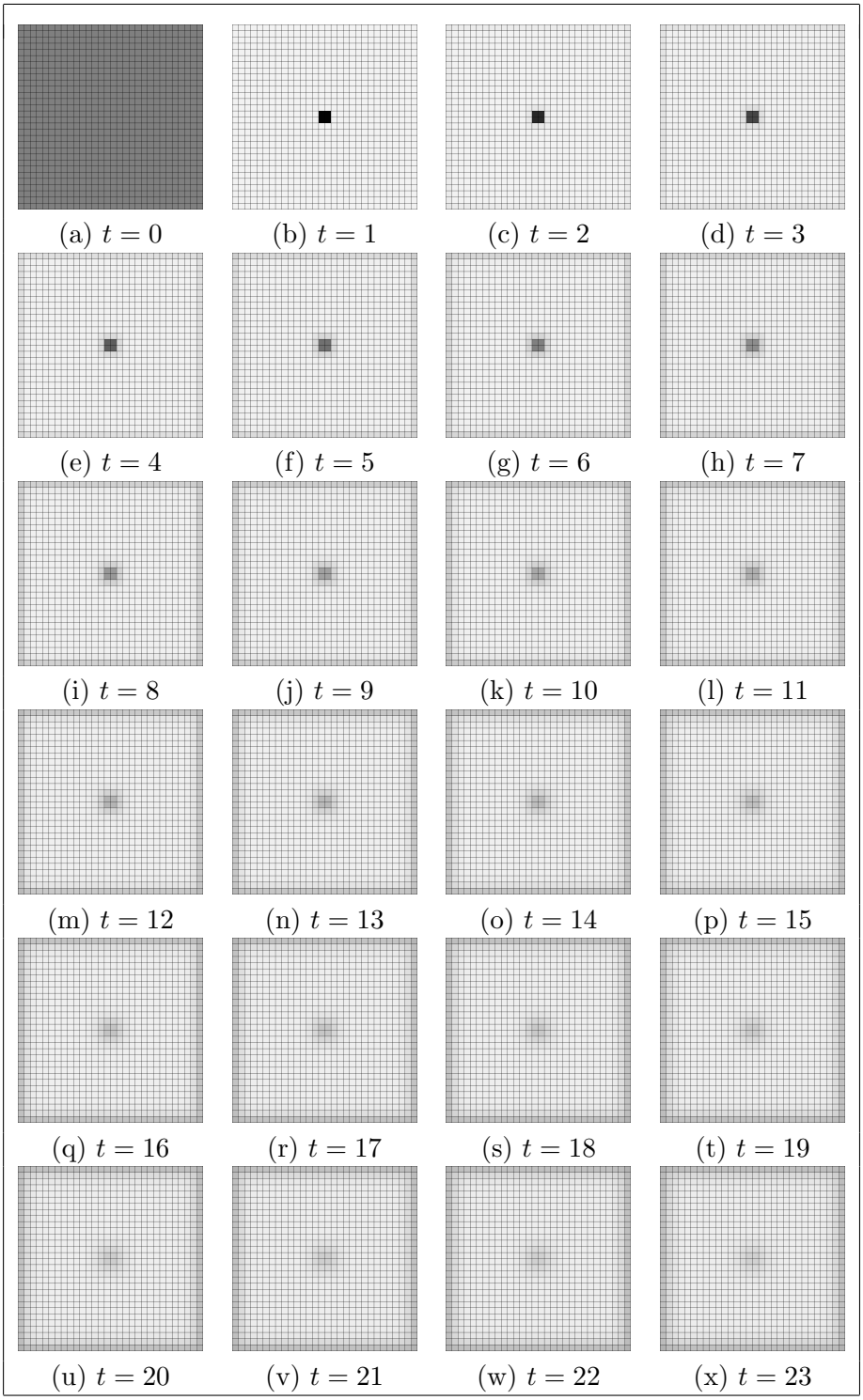


Figura 4.3: Exemplo da etapa de predição baseada no movimento das pessoas. A etapa de predição é aplicada repetidamente após um sensor hipotético detectar pessoas no ambiente, que é representado por uma grade semântica. Quanto mais escura a célula, maior é a probabilidade de existir pessoas.

provenientes dos sensores do robô, tais como odômetros, unidades inerciais ou GPS (Sistema de Posicionamento Global). Como as medições destes sensores estão sujeitas a erros, o modelo de movimento também deve ser descrito de forma probabilística, utilizando uma função da distribuição da posição do robô no tempo t que indica a possibilidade do robô estar em diferentes posições com determinada probabilidade. Consequentemente, a estimativa das posições das pessoas em relação ao sistema de coordenadas do robô estão sujeitas às mesmas incertezas. Para calcular as novas posições em relação ao sistema de coordenadas dos sensores após seus movimentos, o vetor de deslocamento do robô pode ser subtraído das coordenadas das pessoas. Dessa forma, mesmo que nenhuma das pessoas tenha se movido no ambiente, ocorre um movimento relativo quando os sensores se deslocam.

Considerando que a predição baseada no movimento das pessoas é calculada usando a Equação (4.4), a influência do movimento dos sensores é dada por:

$$\overline{bel}(x_t = \mathrm{p}) = P(\mathrm{p}|v_t^{\mathrm{sensores}})\overline{bel}'(x_t = \mathrm{p})\,, \tag{4.8}$$

que é uma predição da presença de pessoas com base na velocidade dos sensores combinada com a velocidade das pessoas no tempo atual e a probabilidade *a posteriori* do estado anterior, antes de incorporar a medição dos detectores de pessoas no tempo t. A Equação (4.9) pode ser reescrita como:

$$\overline{bel}(x_t = \mathrm{p}) = P(\mathrm{p}|z_{t-1}, v_t^{\mathrm{pessoas}}, v_t^{\mathrm{sensores}})\,. \tag{4.9}$$

A probabilidade de pessoas estarem localizadas na região centrada em (a, b) no tempo t baseada no movimento dos sensores é calculada considerando a probabilidade de pessoas de outras regiões terem se movido para (a, b) em relação ao sistema de coordenadas dos sensores. Portanto, o mo-

vimento relativo das pessoas faz com que uma região possa receber pessoas vindas de regiões vizinhas e, como há incerteza nesse movimento, as probabilidades de pessoas de outras regiões moverem-se para a região de interesse são combinadas de acordo com a Equação (4.6). Dessa forma, para calcular probabilidade $P(\mathrm{p}|v_t^{\mathrm{sensores}})$, utiliza-se Equação (4.6), mas substituindo v_t^{pessoas} por v_t^{sensores}. A probabilidade p_k é calculada pelo modelo de movimento dos sensores e está relacionada à probabilidade do movimento relativo de pessoas de (i, j) para (a, b).

O modelo de movimento do sensor utilizado nesta etapa de predição tem como dados de entrada a velocidade do robô, sua posição s_{t-1} no tempo $t-1$ e sua posição hipotética s_t no tempo t. A saída do modelo é a probabilidade do sensor ter se deslocado de uma posição para outra, que é utilizada para calcular a influência do movimento do robô na predição da posição das pessoas (Equação (4.6)). Esse modelo foi adaptado do movimento baseado em odometria de Thrun et al. [2005], que utiliza a velocidade fornecida pelo odômetro do robô, a qual é disponibilizada após o movimento ter sido realizado pelo do robô.

O modelo de movimento utiliza informação de movimento relativo, de uma posição s_{t-1} a outra posição s_t. A velocidade é utilizada para calcular a diferença entre a posição estimada $\overline{s}_t$ no tempo t a partir da posição atual estimada $\overline{s}_{t-1}$, considerando a velocidade linear e a velocidade angular do robô. A diferença entre $\overline{s}_t$ e $\overline{s}_{t-1}$ é usada como estimativa da diferença entre as posições reais s_t e s_{t-1}. No modelo probabilístico de Thrun et al. [2005], que assume que o robô pode ser controlado por meio das velocidades rotacional e translacional e possui propriedades não holonômicas, o movimento relativo entre s_t e s_{t-1} é dado por uma sequência de transformações a partir da posição s_{t-1}, dadas por uma rotação seguida de uma translação e finalmente uma

última rotação, sendo esses três parâmetros independentes e sujeitos a ruído. São calculados pelas equações a seguir, dado que a posição representada por s contém as coordenadas (a, b):

$$\hat{rot1} = atan2(b_t - b_{t-1}, a_t - a_{t-1}) - \theta_{t-1}\,, \tag{4.10}$$

$$\hat{trans} = \sqrt{(b_{t-1} - b_t)^2 + (a_t - a_{t-1})^2}\,, \text{e} \tag{4.11}$$

$$\hat{rot2} = \theta_t - \theta_{t-1} - \hat{rot1}\,, \tag{4.12}$$

onde θ é a direção inicialmente apontada pelo robô e $atan2(c, d)$ é uma função que calcula o arco-tangente de c/d cujo resultado está no intervalo $[-\pi, \pi]$. As rotações e translações obtidas a partir da velocidade são dadas por:

$$rot1 = v^{\text{angular}}\Delta t\,, \tag{4.13}$$

$$trans = v^{\text{linear}}\Delta t\,, \text{e} \tag{4.14}$$

$$rot2 = 0\,, \tag{4.15}$$

onde v^{angular} é a velocidade angular medida, v^{linear} é a velocidade linear medida e Δt é o intervalo de tempo considerado. A rotação final $rot2$ é configurada para zero pois considera-se que a direção final é a mesma da direção do movimento, ou seja, o robô não rotaciona no final. Para simplificar o modelo, assume-se que o robô movimenta-se em um ambiente plano e será utilizado um intervalo de tempo Δt pequeno o suficiente para que seja possível aproximar a velocidade real por uma velocidade constante em cada intervalo de tempo.

Para o cálculo da probabilidade p_k, multiplicam-se as probabilidades dos erros dos três parâmetros $p1$, $p2$ e $p3$: $p_k = p1\ p2\ p3$. As probabilidades dos erros são calculadas pelas diferenças entre os respectivos pares, de acordo com as equações a seguir:

$$p1 = prob(rot1 - \hat{rot1}, v1)\,, \tag{4.16}$$

$$p2 = prob(trans - \hat{trans}, v2)\,, \text{e} \tag{4.17}$$

$$p3 = prob(rot2 - \hat{rot2}, v3)\,, \tag{4.18}$$

onde $prob(c, d)$ é uma função que modela o erro do movimento e calcula a probabilidade de c utilizando uma distribuição normal de média zero e variância d. Os valores da variância são escolhidas de acordo com as características do movimento do robô (por exemplo, maior ou menor erro translacional ou rotacional implica em valores distintos para o parâmetro). Na implementação da metodologia proposta, esses valores foram obtidos empiricamente por meio de simulação similar à da Figura 4.4 e são dados por:

$$v1 = (\alpha_1 rot1^2 + \alpha_2 trans^2)^2 \times 10^6\,, \tag{4.19}$$

$$v2 = (\alpha_3 trans^2 + \alpha_4 rot1^2 + \alpha_4 rot2^2)^2 \times 10^6\,, \text{e} \tag{4.20}$$

$$v3 = \alpha_1 rot2^2 + \alpha_2 trans^2\,, \tag{4.21}$$

onde α_1 e α_2 são iguais a $0,001$ e α_3 e α_4 iguais a $0,0005$.

A Figura 4.4 mostra um exemplo da etapa de predição baseada no movimento dos sensores aplicada repetidamente após um sensor hipotético detectar pessoas no ambiente, que é representado por uma grade semântica com 30×30 células retangulares, cada uma medindo $0,30\,\text{m}$. Inicialmente, um sensor hipotético fixado em um robô móvel observa a presença de pessoas na região central do ambiente representado pela grade (Figura 4.4(a)). A probabilidade de existir pessoas em cada célula determina a sua cor. Quanto mais escura a célula, maior é a probabilidade de existir pessoa. Nos próximos instantes de tempo, o robô inicia um movimento retilíneo uniforme na direção de $\pi/2$ radianos em relação ao referencial do sensor (ou seja, de baixo

para cima na grade) com velocidade linear igual a 2 m/s. As leituras do sensor não serão consideradas para que se possa observar o efeito da predição nas probabilidades. A partir da Figura 4.4(b), em cada instante de tempo é aplicada a Equação (4.9) às probabilidades de cada célula, considerando como crença *a priori* as probabilidades do instante anterior. Observando-se a sequência de imagens da figura, nota-se que a predição baseada no movimento dos sensores tem o efeito de deslocar visualmente as probabilidades das células na direção contrária ao movimento. Como não há leitura de nenhum sensor, ocorre também uma suavização das probabilidades, refletindo a incerteza introduzida pelo movimento do robô. À medida em que o tempo passa, maior é a incerteza da existência de pessoas pois a grade avança para regiões que anteriormente estavam fora dos limites da grade e que não foram observadas pelo sensor (nas regiões fora dos limites da grade a probabilidade de existir pessoas é igual a $0,5$).

A próxima seção descreve a etapa de atualização da metodologia.

4.3 Atualização

A atualização é uma etapa que segue a predição e consiste em determinar a probabilidade de que uma dada região do espaço é ocupada por pessoas usando informações dos detectores de pessoas e também a probabilidade calculada na etapa de predição. Baseada na segunda equação do Filtro de Bayes (Equação (2.6)), esta etapa é também conhecida como correção, pois incorpora uma nova medição z_t à crença, multiplicando $\overline{bel}(x_t)$ pela probabilidade da medição z_t ter sido observada:

$$bel(x_t = \mathrm{p}) = \eta \; P(z_t|x_t) \; \overline{bel}(x_t = \mathrm{p}) \,, \tag{4.22}$$

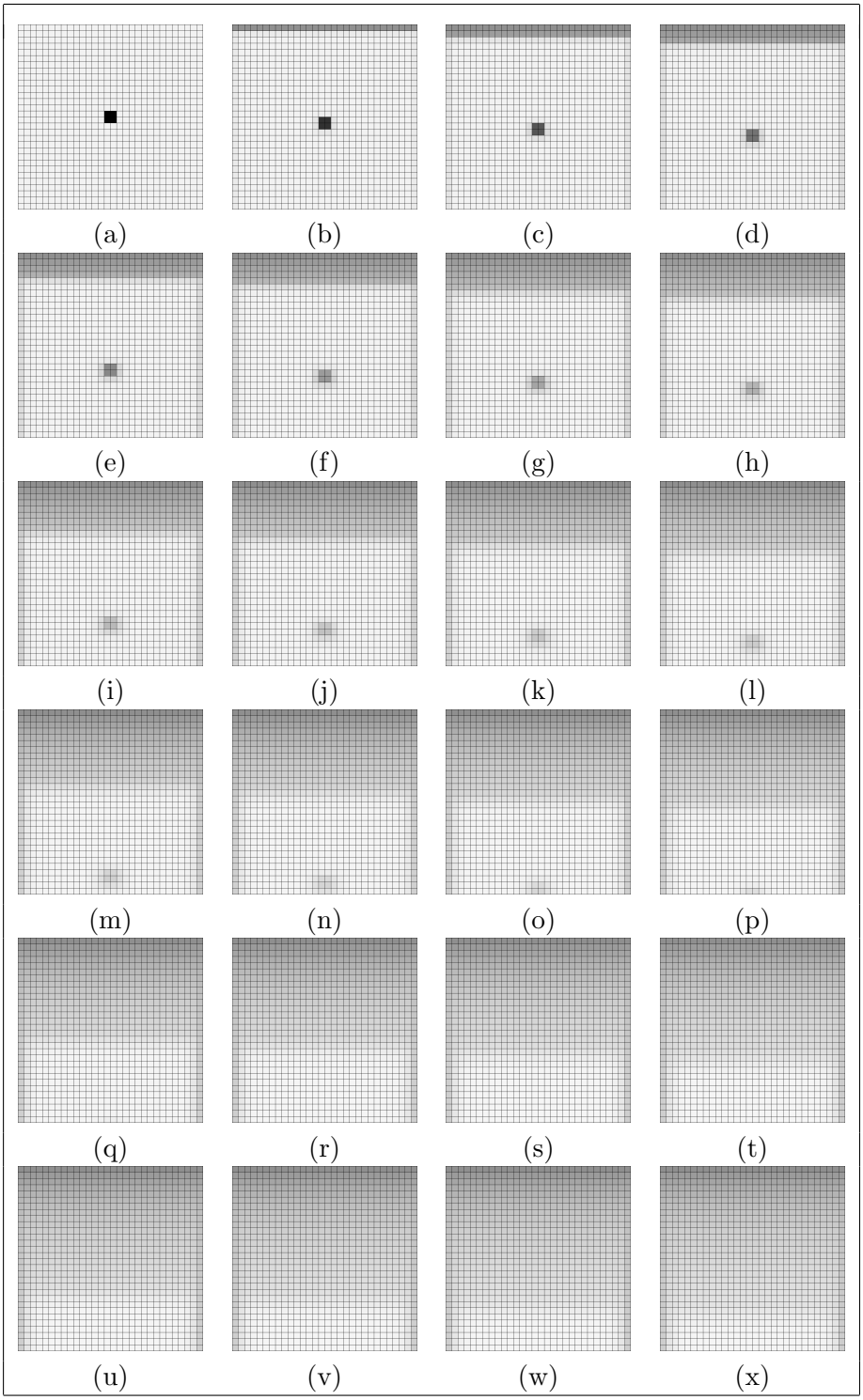


Figura 4.4: Exemplo da etapa de predição baseada no movimento dos sensores. A etapa de predição é aplicada repetidamente após um sensor hipotético detectar pessoas no ambiente, que é representado por uma grade semântica. Quanto mais escura a célula, maior é a probabilidade de existir pessoas.

onde o fator de normalização é dado por $\eta = 1/(bel(x_t = \text{p}) + bel(x_t = \text{np}))$ e $P(z_t|x_t)$ é a a probabilidade do resultado da fusão dos detectores, dado o estado x_t, ou seja, dado um conjunto de resultados de detectores de pessoas $D_1, ..., D_n$ no tempo t, z_t é o resultado da fusão desses detectores.

A metodologia proposta utiliza informações de múltiplos detectores de pessoas na etapa de atualização de forma que os dados dos sensores são processados para que cada detector forneça a posição das pessoas detectadas e uma medida da confiança desta estimativa. Esta informação de alto-nível é combinada de modo que, quando mais de um detector indica a presença de uma pessoa em uma dada posição, a confiança da detecção é maior que a confiança no caso de apenas um detector detectar uma determinada pessoa.

Quando mais de um detector é usado, suas informações devem ser integradas levando-se em consideração as diferentes características dos detectores. Para combinar múltiplos sensores, existem diversas possibilidades na literatura da área de fusão sensorial. Entre elas, pode-se citar as seguintes funções de combinação:

- Uma média ponderada das distribuições de probabilidade das observações dos sensores (*Linear Opinion Pools*) [Adarve et al., 2012];

- Uma função que combina PMFs (função massa de probabilidade) baseada na teoria de conjuntos, que representam a ocorrência de estados determinados, como por exemplo livre, ocupado, desconhecido e resultante de um conflito (Dempster-Shafer) [Moras et al., 2011];

- Uma função que utiliza a Regra de Bayes para combinar as probabilidades condicionais (dado um estado x, a probabilidade de uma medição ter sido observada) e obter a probabilidade *a posteriori* [Yguel et al., 2006]; ou

- Uma função para calcular a probabilidade máxima entre as células correspondentes [Thrun et al., 2005].

Como descrito em Baig et al. [2014], informações conflituosas podem gerar inconsistências em alguns desses métodos. Na metodologia proposta é utilizado o conceito de fusão de dados de múltiplos sensores para combinar de maneira apropriada as informações de alto-nível de detectores de pessoas. Para lidar com informações conflitantes e reduzir a confiança daqueles detectores que não fornecem informações relevantes ao processo, nesta tese os dados de todos os detectores aplicados aos dados de um ou mais sensores distintos são combinados utilizando a Lei de De Morgan [Thrun et al., 2005]. Considerando que existem N detectores cujos resultados são $D_1, D_2, D_3 \ldots D_N$, a confiança da fusão quando pelo menos um dos detectores detecta pessoas em uma dada região do espaço pode ser calculada por:

$$P(z_t|x_t = p) = \ 1 - \prod_{i=1:N} \left(1 - P(D_i|x_t = \mathrm{p})\right). \tag{4.23}$$

A probabilidade do resultado da fusão dos detectores ser pessoa depende do fato de algum dos detectores ter detectado pessoa ou não. Quando um ou mais detectores detectam pessoa, utiliza-se a regra de De Morgan pois essa regra faz com que a confiança seja aumentada quando mais de um detector encontram pessoas.

Para calcular $P(D_i|x_t = \mathrm{p})$ são usadas as métricas de Valor Preditivo Positivo (VPP) e Taxa de Falsos Negativos (TFN). O VPP, também conhecido como precisão, é a razão entre o número de detecções corretas e o número total de detecções. É usado para medir o número de acertos em relação ao número total de objetos identificados pelo classificador como positivos. No caso deste trabalho, o VPP é usado como uma aproximação da confiança do detector quando este detecta pessoas e portanto $P(D_i = p|x_t = \mathrm{p}) = \mathrm{VPP}$.

Quando o detector não indica a presença de pessoas, a confiança é aproximada por $P(D_i = np|x_t = \text{p}) = \text{TFN}$, que é a razão entre o número de falsos negativos e o número total de pessoas presentes na cena e está associada à probabilidade do detector não detectar uma pessoa quando havia pessoa na cena. Os valores de VPP e TFN podem ser obtidos experimentalmente para cada detector.

Quando nenhum dos detectores encontra pessoas em determinada região, a Lei de De Morgan não é utilizada, pois deseja-se que o modelo reduza a probabilidade de existir pessoas na região. Para garantir que a probabilidade de existir pessoas é reduzida, a confiança associada à presença de pessoas nessa região é igual à do sensor mais confiável em detectar as pessoas presentes na cena, ou seja, com a menor TFN dentre os detectores. Consequentemente, a probabilidade *a posteriori* de ter pessoa é minimizada. Logo:

$$P(z_t|x_t = \text{p}) = \min_i \left(P(D_1 = \text{np}|x_t = \text{p}),\ \ldots\ ,\ P(D_N = \text{np}|x_t = \text{p})\right). \quad (4.24)$$

onde $\min_i$ é uma função que calcula o menor valor $P(D_i = \text{np}|x_t = \text{p})$ de todos os N detectores.

O Algoritmo 1 resume as etapas principais da metodologia. A próxima seção explica a utilização de grades de ocupação semânticas na implementação das etapas da metodologia de forma a prover informações sobre a presença de pessoas em cada região do ambiente e uma medida de confiança dessa informação.

4.4 Implementação da metodologia proposta

A metodologia proposta pode ser implementada utilizando técnicas em que o cálculo da probabilidade *a posteriori* esteja associado a regiões do es-

paço. Neste caso, a representação do espaço deve ser discretizada por exemplo em partículas ou células. A técnica que possibilita o uso de partículas é o Filtro de Partículas, que representa a distribuição *a posteriori* por um conjunto de amostras ponderadas. A ideia do Filtro de Partículas é manter um conjunto de N partículas, formadas por amostras dos estados e seus pesos. Quando uma nova medição acontece, o peso da partícula é recalculado como a probabilidade da observação ter ocorrido dado o estado da partícula. As partículas são então reamostradas de acordo com o seu peso [Gidel et al., 2009; Choset et al., 2005]. Aplicando-se ao problema de detecção de pessoas, uma grande concentração de partículas em determinada região do espaço indicaria uma probabilidade proeminente de existir pessoas nesta região. Embora seja conhecido como uma técnica de baixo custo computacional, a desvantagem é que podem haver regiões do espaço com probabilidade de existir pessoas que não vão estar relacionadas a partículas, a não ser que o número de partículas seja grande o suficiente, o que pode aumentar o custo computacional da técnica.

A representação do espaço por células tem seu fundamento nas grades de ocupação, como explicado na Seção 2.3. Para a implementação da metodologia proposta foi utilizada uma grade de ocupação semântica, que é uma representação espacial do ambiente derivada das grades de ocupação, porém contendo as posições de objetos de determinadas classes [Nüchter e Hertzberg, 2008]. As grades de ocupação semânticas são empregadas em diversas aplicações de robótica [Liu e von Wichert, 2014; Adarve et al., 2012] por serem uma representação compacta do ambiente ao redor do robô. Além de ser possível prover informações sobre a ocupação de espaço, a grade de ocupação semântica permite a representação de outros tipos de objetos detectados. Outra vantagem é que a grade reduz o problema de associação de dados de

Entrada: O conjunto $B' = \{bel(x_{t-1}^{i,j}) | (i,j) \in G\}$ de probabilidades *a priori* das regiões do espaço em relação à presença de pessoas no instante de tempo $t-1$; modelo de movimento das pessoas; modelo de movimento dos sensores; velocidade dos sensores v_t^{sensores}; e resultados dos detectores $D_1, ..., D_n$ em t.

Saída: O conjunto $B = \{bel(x_t^{i,j}) | (i,j) \in G\}$ de probabilidades *a posteriori* para cada região do espaço no instante de tempo t.

início

para *cada região centrada em* (i,j) *tal que* $(i,j) \in G$ **faça**

//Predição baseada no movimento das pessoas

$\overline{bel}'(x_t^{i,j} = \text{p}) = P(\text{p}|x_{t-1}^{i,j} = \text{p}, v_t^{\text{pessoas}} = 0) bel(x_{t-1}^{i,j} = \text{p}) +$

$+ P(\text{p}|x_{t-1}^{i,j} = \text{np}, v_t^{\text{pessoas}} = v^{\text{média}}) bel(x_{t-1}^{i,j} = \text{np})$

fim

para *cada região centrada em* (i,j) *tal que* $(i,j) \in G$ **faça**

//Predição baseada no movimento dos sensores

$\overline{bel}(x_t^{i,j} = \text{p}) = P(\text{p}|v_t^{\text{sensores}})\ \overline{bel}'(x_t^{i,j} = \text{p})$

fim

para *cada região centrada em* (i,j) *tal que* $(i,j) \in G$ **faça**

Atribua o resultado dos detectores referente à região (i,j) a $z_t^{i,j}$

//Fusão dos resultados dos detectores

se *algum detector indicou a presença de pessoa na região* **então**

$P(z_t^{i,j}|x_t^{i,j} = p) = 1 - \prod_{i=1:N}(1 - P(D_i|x_t^{i,j} = \text{p}))$

senão

$P(z_t^{i,j}|x_t^{i,j} = \text{p}) = \min_i\ (P(D_1 = \text{np}|x_t^{i,j} = \text{p}),\ ...\ ,$

$P(D_N = \text{np}|x_t^{i,j} = \text{p}))$

fim

//Atualização

$bel(x_t^{i,j} = \text{p}) = \eta\ P(z_t^{i,j}|x_t^{i,j} = \text{p})\ \overline{bel}(x_t^{i,j} = \text{p})$

fim

fim

Algoritmo 1: Algoritmo que resume as etapas da metodologia proposta para um determinado instante de tempo t.

múltiplos sensores a um simples mapeamento para as suas células [Yoder et al., 2014]. A próxima subseção explica como foram utilizadas grades de ocupação semânticas na implementação da metodologia.

4.4.1 Grades de ocupação semânticas

A implementação da metodologia proposta consiste em particionar o ambiente em uma grade bidimensional de células retangulares e, utilizando informações de alto nível proveniente do processamento dos dados dos diversos detectores, atribuir a cada célula uma probabilidade de a área representada pela célula ser ocupada por pessoas. Um exemplo de grade semântica é mostrado na Figura 4.5. A Figura 4.5(a) mostra uma cena vista de cima contendo três pessoas e um objeto sendo escaneados por um sensor a laser, cujos raios estão representados pelas linhas azuis, posicionado no centro da borda inferior da figura. A Figura 4.5(b) mostra uma grade semântica com 900 (30×30) células que representa a cena em (a). As células da grade estão coloridas de acordo com as leituras do laser: branco para células livres, vermelho para células ocupadas por pessoas, preto para células ocupadas por outros obstáculos e cinza para células fora do alcance do sensor.

A metodologia foi implementada utilizando uma grade semântica com 30×30 células quadradas, cada uma com os lados medindo $0,3\,\mathrm{m}$. As dimensões das células permitem que uma pessoa ocupe mais de uma célula, bem como duas ou mais pessoas dividam a mesma célula. Essas dimensões foram escolhidas tendo como referência o modelo do movimento das pessoas (Seção 4.1), que considera que pessoas existentes no ambiente ocupam regiões distintas e portanto a célula possui uma área próxima do espaço ocupado por uma pessoa. A quantidade de células abrange uma área de $9 \times 9\,\mathrm{m}$, que é razoável para aplicações de robótica em que os robôs possuem velocidades

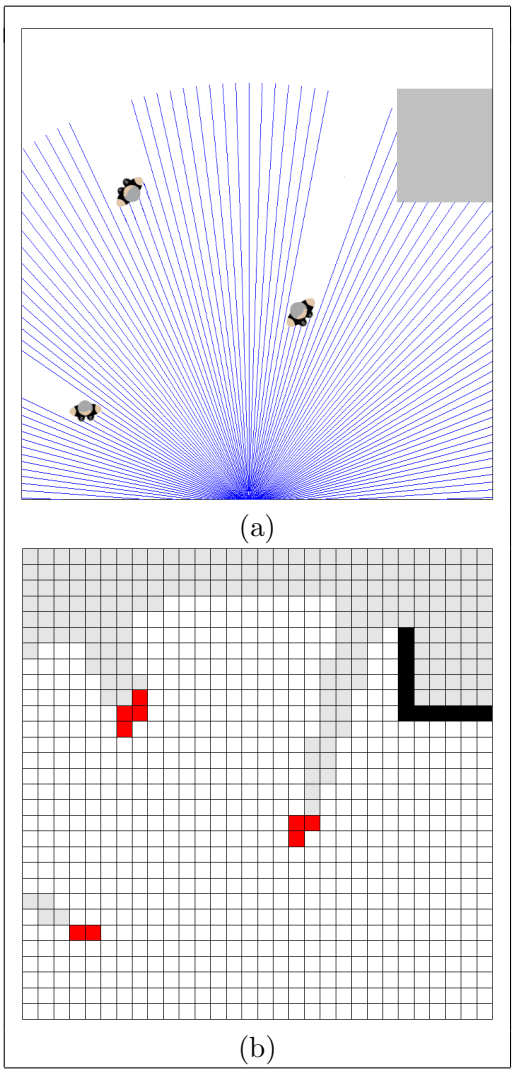


Figura 4.5: Exemplo de grade de ocupação semântica. (a) Uma cena contendo pessoas e um outro objeto sendo escaneado por um sensor a laser posicionado no centro da borda inferior da figura. (b) A grade semântica que representa a cena com células coloridas de acordo com a informação obtida pelas leituras do sensor.

comparáveis à velocidade das pessoas. Quanto maior a velocidade do robô, maior deverá ser o tamanho da grade (e também o campo de visão dos sensores) para uma navegação segura que considere uma distância para frenagem. Para aplicações em veículos inteligentes, por exemplo, devem ser consideradas as distâncias de acordo com o veículo utilizado, como pode ser visto na Tabela 4.1, que mostra as distâncias para frenagem para veículos de passeio atualmente no mercado [Messeder, 2014].

A grade semântica implementada está posicionada no mesmo plano dos raios do laser de forma similar à situação ilustrada na Figura 4.5(a). Os detectores fornecem as posições das pessoas detectadas, que são mapeadas para a grade. Como os sensores tem limitações e não estão livres de ruído, existe uma incerteza associada tanto à localização fornecida quanto ao resultado da detecção. Nesta implementação, são considerados os sensores laser e câmera monocular. A incerteza na posição fornecida pelo laser foi desconsiderada por ser desprezível em comparação às dimensões de uma pessoa (o erro na posição dos objetos detectados é de aproximadamente 35 mm de acordo com SICK [2006]). Neste caso, cada ponto atingido por raio do laser que foi classificado como pessoa pelo detector é mapeado para a célula que contém as coordenadas do ponto transformadas para o referencial da grade. A Figura 4.6 mostra exemplos de dados do laser plotados em um gráfico e a grade correspondente.

Em relação ao detector de pessoas em imagens, como está sendo considerada uma câmera monocular, utiliza-se o método proposto por Stein et al. [2003] para estimar a distância entre a câmera e a pessoa detectada. Este método é baseado em uma abordagem que assume que a altura da câmera em relação ao solo é conhecida. A incerteza desta estimativa foi obtida experimentalmente com a utilização de imagens com pessoas em diferentes posições

Velocidade	Distância para frenagem
60 km/h	13,6 a 15,9 m
80 km/h	24,5 a 28,7 m
100 km/h	40,0 a 45,9 m

Tabela 4.1: Distância para frenagem de veículos obtidos em teste na mídia (modelos Ka+, Logan, Prisma e Grand Siena) [Messeder, 2014].

e suas respectivas distâncias medidas por um laser. A diferença entre a distância medida pelo laser e a distância obtida pelo método de Stein et al. [2003] foi calculada para cada imagem e o erro da estimativa foi representado por uma PDF aproximada pela distribuição normal com média 0 e desvio padrão de 1.0 m no eixo x e 1.7 m no eixo y, ambos no referencial da câmera. A Figura 4.7 mostra o efeito da distribuição na localização de pessoas detectadas pela câmera nas células $(15, 15)$ e $(15, 16)$ da grade.

A implementação da metodologia proposta descrita nesta seção tem como resultado final uma grade semântica com as probabilidades *a posteriori* de cada célula estar ocupada por pessoas. Em aplicações reais, a grade final deve ser processada em uma etapa posterior de classificação para determinar as células que realmente contenham pessoas com base nas probabilidades resultantes e, no caso de terem sido detectadas, determinar também qual a sua localização. Nesta tese, esta etapa de pós-processamento é chamada de classificação (Figura 4.1). Vários métodos de classificação podem ser usados, incluindo a aplicação de um simples limiar em que assume-se que existem pessoas nas células cujas probabilidades correspondentes sejam maiores que um dado valor. Métodos mais sofisticados de classificação também podem ser usados, por exemplo, levando-se em consideração determinadas propriedades das células tais como área, velocidade, dentre outras, para formar regiões similares. Neste contexto, essas regiões são chamadas de *blobs* e a classificação

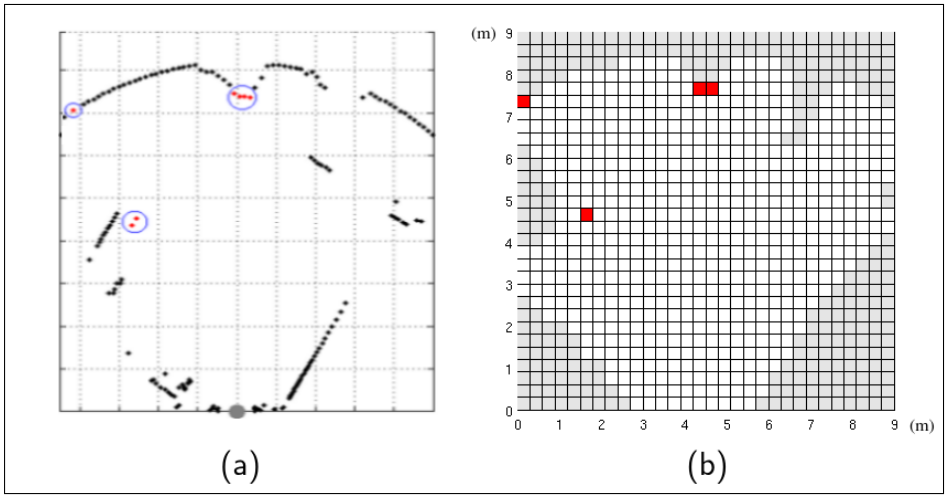


Figura 4.6: Dados da detecção do laser projetado em uma grade com 30×30 células. A detecção de pessoas foi realizada utilizando-se o algoritmo de Bellotto e Hu [2009]. (a) Os dados de um laser 2D estão plotados em um gráfico (cada unidade equivale a 1 metro) e as pessoas detectadas são destacadas pelas circunferências. A posição do laser é marcada por um círculo cinza na parte inferior do gráfico. (b) Desenho da grade de ocupação semântica com informações sobre as pessoas na cena de acordo com o detector do laser (as células vermelhas representam pessoas, as células cinza estão fora do alcance do sensor e as células brancas representam espaço livre).

de *blobs* refere-se à seleção de regiões que possuem propriedades diferentes em relação à regiões vizinhas e sua classificação como pessoas e não pessoas. A classificação de *blobs* na grade de probabilidades *a posteriori* quantifica as pessoas detectadas e permite seu uso em aplicações. Na subseção a seguir, a classificação de *blobs* será explicada.

4.4.2 Classificação de *blobs*

O algoritmo de classificação de *blobs* proposto é baseado na similaridade de probabilidades. O primeiro passo do algoritmo é a criação dos *blobs*, que começa com a seleção das células cujo valor de probabilidade é maior que um determinado limiar (o valor $0,26$ foi utilizado na implementação).

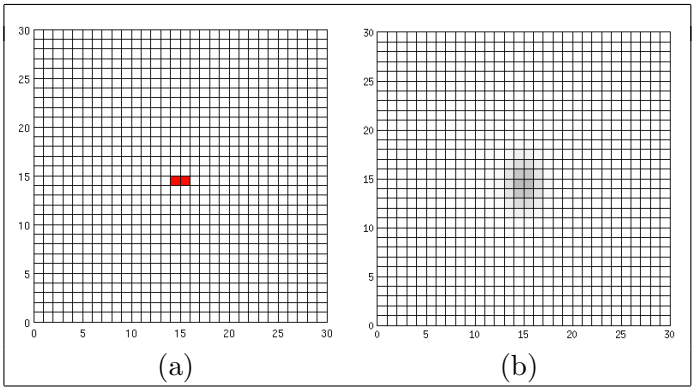
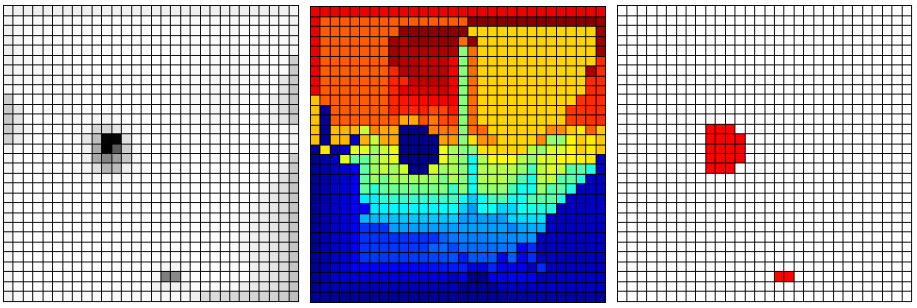


Figura 4.7: Efeito da distribuição na localização de pessoas. (a) As células vermelhas representam as pessoas encontradas pelo detector. (b) Localização das pessoas detectadas em (a) após a aplicação da distribuição do erro. Nesta figura, quanto mais escura for a célula, maior será a probabilidade de existir pessoas.

Células selecionadas que são vizinhas passam a fazer parte do mesmo *blob*. No próximo passo, todas as células que não pertencem a um *blob*, mas que são vizinhas a um *blob*, são visitadas. Para cada célula visitada, verifica-se se a diferença entre a probabilidade da célula e das células do *blob* vizinho é menor que um limiar (o valor utilizado para o limiar é igual a $0,40$) e então a célula é também incluída no *blob*. Caso contrário, um novo *blob* é criado a partir da célula visitada. Este procedimento continua até que todas as células visitadas façam parte de um *blob*. A Figura 4.8(b) mostra a seleção de *blobs* aplicada à grade de atualização da Figura 4.8(a).

O último passo da classificação de *blobs* é a binarização da grade contendo os *blobs*, permitindo a separação entre os *blobs* que contêm pessoas daqueles que não contêm. Os *blobs* recebem o valor 1, que indica a presença de pessoas, se todas as células do *blob* tiverem um valor de probabilidade maior que $0,1$, se possuem um valor de probabilidade de alguma célula maior que a média dos *blobs* da vizinhança e se não possuem tamanho maior que 3 células. Se o *blob* tiver mais que 3 células, ele só será selecionado se passar pelos seguintes



(a) (b) (c)

Figura 4.8: Classificação aplicada à grade de probabilidades *a posteriori*. (a) Grade de probabilidades *a posteriori*, onde quanto mais escura é a célula, maior é a probabilidade de existir pessoas. (b) Seleção dos *blobs*. (c) Binarização da grade em (b), em que as pessoas encontradas estão posicionadas nas células vermelhas.

critérios:

- se possui maior valor que vizinhança em alguma de suas células;
- se não tem células nas bordas da grade;
- se tiver um número de linhas ou colunas menor que 8; e
- se a razão do número de linhas e colunas for maior ou igual a $0,5$.

O Algoritmo 2 apresenta em mais detalhes a binarização da grade de probabilidades *a posteriori*. Todos os parâmetros foram ajustados empiricamente utilizando uma base de dados de treino que será apresentada na Subseção 5.1.2, sendo específicos para esta base. Um exemplo desse algoritmo aplicado à grade de probabilidades *a posteriori* gerada como resultado da metodologia proposta é mostrado na Figura 4.8(c).

No próximo capítulo são apresentados os resultados experimentais utilizando a metodologia proposta e um exemplo de aplicação. Nos experimentos com a implementação descrita, quatro detectores distintos que obtém dados de dois sensores fixos em um robô móvel são usados para validar a metodologia.

Entrada: Grade contendo os *blobs*.

Saída: Grade com valores binários.

início

- **para** *cada blob da grade* **faça**
  - **se** *(média de probabilidades do blob)* $\leq 0,05$ *OU (maior valor das células do blob)* $< 0,1$ **então**
    - Atribui 0 às células do *blob*;
  - **senão**
    - **se** *(total de células do blob)* ≤ 3 **então**
      - **se** *(blob possui célula com maior valor que vizinhança)* **então**
        - Atribui 1 às células do *blob*;
      - **senão**
        - Atribui 0 às células do *blob*;
      - **fim**
    - **senão**
      - **se** *(número células nas bordas da grade)* ≤ 3 **então**
        - Atribui 0 às células do *blob*;
      - **senão**
        - **se** *(blob possui célula com maior valor que vizinhança) E (número de células nas bordas da grade)* $= 0$ *E (máximo de linhas)* < 8 *E (máximo colunas)* < 8 *E (razão entre linhas e colunas)* $\geq 0,5$ **então**
          - Atribui 1 às células do *blob*;
        - **senão**
          - Atribui 0 às células do *blob*;
        - **fim**
      - **fim**
    - **fim**
  - **fim**

fim

Algoritmo 2: Algoritmo de binarização da grade de probabilidades *a posteriori*.

Capítulo 5

Resultados experimentais

Este capítulo apresenta os experimentos realizados em bases de dados reais. A metodologia proposta, implementada utilizando grades de ocupação semânticas, é avaliada qualitativamente e quantitativamente, mostrando ser uma estratégia robusta de fusão de detectores e que pode ser utilizada em diversas aplicações, como por exemplo no rastreamento de pessoas.

A próxima seção descreve o arcabouço experimental, detalhando quais os sensores utilizados e como foram calibrados, bem como a obtenção das bases de dados.

5.1 Arcabouço experimental

Os experimentos foram realizados em duas etapas: experimentos preliminares com detectores individuais e experimentos com a fusão dos detectores. Na primeira etapa, foram escolhidos detectores de pessoas para testes individualmente, sem a utilização da metodologia proposta, o que permitiu conhecer o comportamento dos detectores e obter indicativos sobre a forma mais adequada para a fusão. Este foi um experimento preliminar, utilizando os dados de um laser e uma câmera sob um suporte fixo em ambiente interno, como mostrado na Figura 5.1. O laser utilizado foi o SICK LMS291-S05, que possui ângulo de visão de 180°, frequência de 9 Hz, alcance configurado para

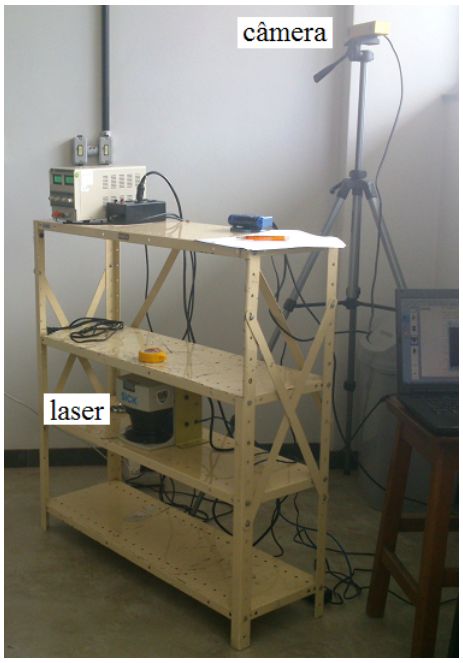


Figura 5.1: Sensores utilizados nos experimentos preliminares.

aproximadamente 8 metros e resolução angular de $0,5°$. A câmera utilizada foi a Bumblebee2 da Point Grey Research. Embora seja uma câmera estéreo, neste trabalho foram utilizadas apenas as imagens geradas pela câmera esquerda, a uma frequência de 7 Hz. Cada imagem gerada tem tamanho 640×480 pixeis.

O laser e câmera estavam posicionados a aproximadamente 40 cm e 150 cm do chão, respectivamente. O laser estava aproximadamente alinhado à câmera em relação ao eixo x e deslocado de 70 cm à frente. A rotação da câmera em relação ao laser fornecido pela calibração (explicada na próxima subseção) é dada pelos ângulos de rotação em torno dos eixos fixos x, y e z do seguinte vetor (em radianos):

$$R_1 = \{0,314;\ -0,384;\ -0,069\}.$$

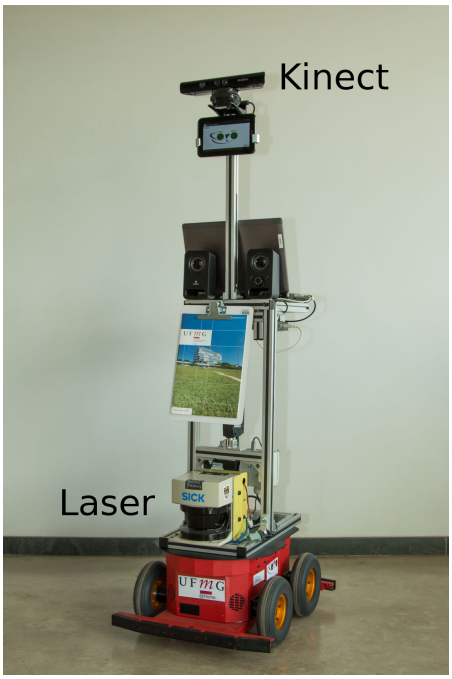


Figura 5.2: Plataforma robótica do CORO/UFMG chamada MARIA (*Manipulator Robot for Interaction and Assistance*).

Na segunda etapa dos experimentos, com o objetivo de validar a metodologia proposta, foi utilizada uma plataforma robótica móvel locomovendo-se autonomamente em um ambiente interno usando a metodologia proposta por Araujo et al. [2015]. Esta plataforma, chamada MARIA (*Manipulator Robot for Interaction and Assistance*), possui um sensor a laser 2D e um sensor Kinect, como mostrado na Figura 5.2.

O primeiro sensor a laser utilizado na MARIA foi o modelo SICK LMS291-S05, fixo a $0,29$ m acima do chão e configurado para alcance máximo de 32 metros, frequência de 9 Hz, ângulo de visão de 180° e resolução angular de 1°. O laser estava aproximadamente alinhado em relação ao eixo x da câmera, porém deslocado cerca de $0,20$ m à frente. A orientação da câmera em relação ao laser obtida pela calibração é dada pelos ângulos em torno dos eixos fixos

x, y, e z do seguinte vetor (em radianos):

$$R_2 = \{0,122;\ -0,003;\ -0,017\}.$$

O segundo sensor a laser utilizado em substituição ao primeiro foi o SICK LMS100, posicionado a $0,31\,\text{m}$ acima do chão e configurado para alcance máximo de 20 metros, frequência de 50 Hz, ângulo de visão de 270° e resolução angular de $0,5°$. Esse laser também estava aproximadamente alinhado em relação ao eixo x da câmera, mas deslocado cerca de $0,20\,\text{m}$ à frente. A orientação da câmera em relação ao segundo laser obtida pela calibração é dada pelos ângulos em torno dos eixos fixos x, y, e z do seguinte vetor (em radianos):

$$R_3 = \{0,201;\ -0,011;\ -0,006\}.$$

O Kinect é um sensor de baixo custo formado por uma câmera RGB além de uma câmera e um projetor de infra-vermelho e estava fixo a $1,63\,\text{m}$ acima do chão. A câmera RGB fornece imagens com 640×480 pixeis e seu campo de visão horizontal é de 62°. A câmera de profundidade também fornece uma imagem 640×480 pixeis e seu campo de visão horizontal é de 58°. A frequência original das câmeras é de 30 Hz, porém foram subamostradas para 10 Hz, reduzindo assim o custo de processamento das imagens.

Para a captura dos dados, utilizou-se o ROS [Quigley et al., 2009], que é um programa que provê bibliotecas e ferramentas para aplicações em robótica. Foram criados nós para processamento dos dados do laser e das câmeras, de forma que é possível utilizá-los tanto de forma aproximadamente sincronizada como também de forma assíncrona, já que a metodologia pro-

posta permite a atualização de informações provenientes do processamento dos sensores a diversas frequências. Como as frequências dos sensores são diferentes, a sincronização, quando necessária nos experimentos, foi realizada tomando-se a leitura do laser temporalmente mais próxima da última imagem fornecida pela câmera. A metodologia proposta foi implementada no Matlab, recebendo os dados dos detectores já processados por meio de arquivos, de modo off-line, pois os experimentos foram realizados após a coleta de dados.

5.1.1 Calibração dos sensores

A calibração é uma etapa essencial do trabalho pois, sem ela, não é possível relacionar as informações de sensores diferentes, já que é necessário conhecer a posição do laser em relação à câmera e vice-versa. Com a calibração obtém-se um mapeamento que transforma pontos no sistema de referência do laser para o sistema de referência da câmera e então para o plano da imagem [Premebida et al., 2009]. A posição relativa dos sensores não foi alterada durante o experimento.

O problema da calibração consiste em estimar os valores dos parâmetros intrínsecos da câmera e também os parâmetros extrínsecos em relação ao referencial do laser. Conforme descrito em [Trucco e Verri, 1998]:

- Parâmetros extrínsecos: são parâmetros que definem a localização e orientação do sistema de coordenadas de referência da câmera em relação a um sistema de coordenadas do mundo conhecido. Tipicamente, a transformação entre os sistemas de coordenadas da câmera e do mundo é descrita por um vetor de translação e uma matriz de rotação, que descrevem a posição relativa das origens dos dois sistemas de coordenadas de referência e a rotação dos eixos.

- Parâmetros intrínsecos: são parâmetros necessários para relacionar as coordenadas em pixeis de um ponto na imagem com as coordenadas correspondentes no sistema de coordenadas de referência da câmera. Para um modelo de câmera escura (câmera *pinhole*[1]) existem três conjuntos de parâmetros intrínsecos: a projeção perspectiva (distância focal), a transformação entre as coordenadas do sistema de coordenadas da câmera e coordenadas em pixeis e a distorção geométrica.

Existem pacotes disponíveis livremente para calibração de câmera e câmera-laser, tais como [Bouguet, 2010] para calibração de câmera, [Zhang e Pless, 2004] para calibração câmera-laser e [Unnikrishnan e Hebert, 2005] para calibração de câmera e laser 3D. Entretanto, esses pacotes requerem bastante esforço do operador para atingir resultados confiáveis, pois parte da calibração envolve marcações manuais. O trabalho de Kassir e Peynot [2010] apresenta um método automático e robusto para calibração de uma câmera com laser 2D, utilizando os métodos de Bouguet [2010] e Zhang e Pless [2004] porém automatizando as etapas de seleção de quinas do padrão de calibração e de marcação dos pontos da leitura do laser referentes ao padrão.

Assim, para a calibração câmera-laser foi escolhido o pacote Matlab disponibilizado por Kassir e Peynot [2010][2]. Inicialmente, é realizada a calibração da câmera, em que os parâmetros intrínsecos da câmera são estimados, bem como a transformação rígida entre a câmera e um padrão de calibração para cada imagem. A transformação é necessária para a calibração câmera-laser, que usa os pontos originados do padrão de calibração que aparecem na leitura

[1]Câmera *pinhole*: possui a abertura da câmera reduzida a um ponto. Isso significa que apenas um raio de qualquer ponto pode entrar na câmera, criando uma correspondência um-para-um entre os pontos visíveis, raios de luz e pontos na imagem [Trucco e Verri, 1998].

[2]Automatic Camera-Laser Calibration: http://www-personal.acfr.usyd.edu.au/akas9185/AutoCalib/.

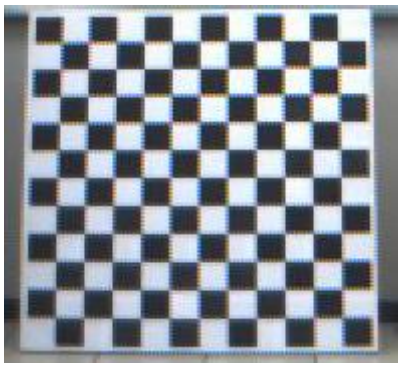


Figura 5.3: Padrão de calibração. O lado dos quadrados brancos e pretos medem 80 mm.

do laser para encontrar a transformação rígida do laser para a câmera. O padrão de calibração utilizado pode ser visto na Figura 5.3 e a Figura 5.4 mostra a detecção da linha do padrão nos dados de leitura do laser.

Como resultado da calibração, obtém-se o vetor de translação e a matriz de rotação representando a orientação e posição da câmera em relação ao laser. Dessa forma, é possível projetar os pontos da leitura do laser na imagem da câmera, como é mostrado na Figura 5.5.

As informações obtidas da calibração também permitem mapear as coordenadas de células da grade de ocupação semântica para a imagem, ou seja, as coordenadas das pessoas no sistema de coordenadas da grade são transformadas em coordenadas no referencial do laser e por fim transformadas em coordenadas da imagem. Um exemplo do mapeamento da grade para a imagem é mostrado na Figura 5.6.

5.1.2 Bases de dados

Três bases de dados foram produzidas para os experimentos, sendo a Base 1 com dados coletados pelos sensores estáticos (Figura 5.1), a Base 2 pelos sensores da MARIA (Figura 5.2) utilizando o laser SICK LMS291-S05

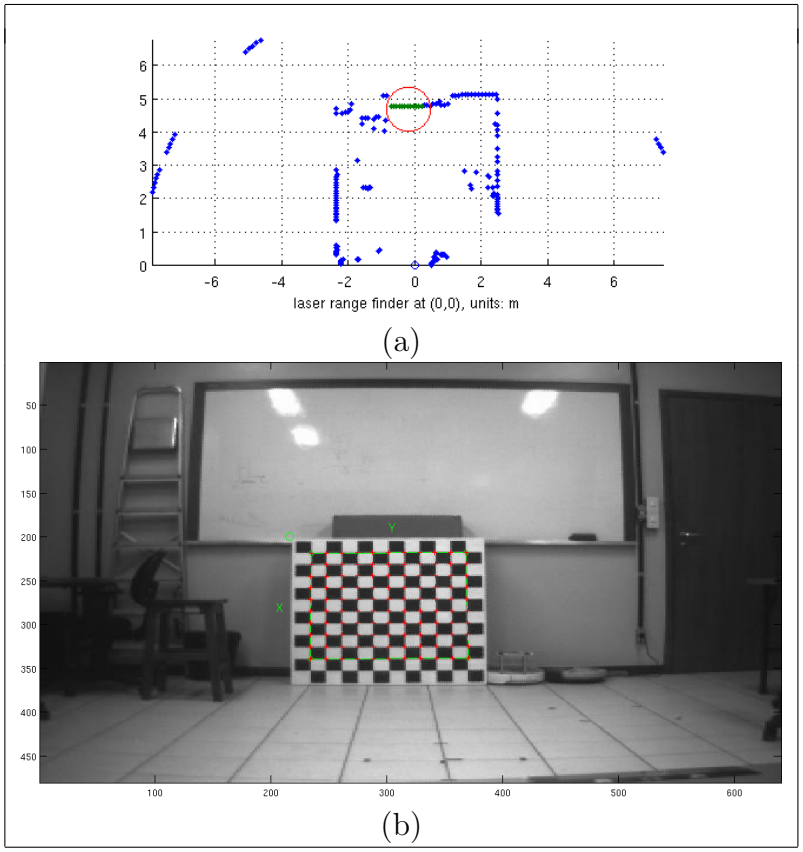


Figura 5.4: Leitura do laser. (a) A detecção da linha referente ao padrão de calibração está destacada com uma circunferência. (b) A leitura do laser em (a) corresponde à posição do padrão no centro da figura, em meio a outros objetos do ambiente.

e as imagens geradas pela câmera RGB do Kinect e a Base 3 pelos sensores da MARIA utilizando o laser SICK LMS100, as imagens geradas pela câmera RGB e os dados de profundidade fornecidos pelo Kinect. As bases possuem características distintas conforme descrito a seguir:

- Base 1: para esta base os sensores foram posicionados no interior do prédio da Escola de Engenharia/UFMG e, como não houve movimento dos sensores, o fundo do ambiente é estático. Durante a obtenção da base, diversas pessoas passaram em frente aos sensores (caminhando,

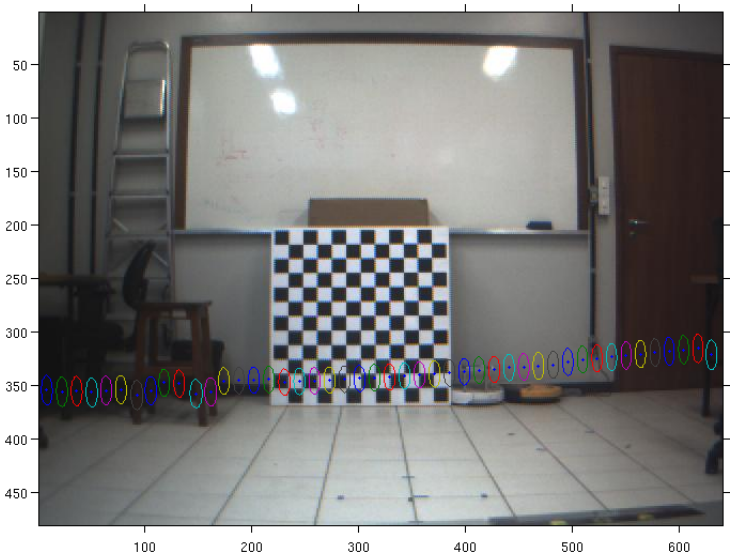


Figura 5.5: Pontos da leitura do laser projetados na imagem da câmera. As elipses coloridas representam a incerteza da transformação.

correndo e também pararam formando grupos de pessoas). Foram gerados 2719 leituras do laser e 1693 imagens.

- Base 2: também foi obtida no interior do prédio da Escola de Engenharia/UFMG, porém o robô no qual os sensores estavam fixos fez um percurso passando por vários corredores abertos e fechados, que, em termos de iluminação, se assemelham a ambientes externos. Esta base permite uma avaliação das diferentes situações presenciadas pelo robô, que se movia variando-se suas velocidades linear e angular e passando por várias pessoas que encontravam-se no ambiente movimentando-se espontaneamente. Foram geradas 6658 leituras do laser e 21029 imagens, sendo que após a subamostragem restaram 6663 imagens. Algumas imagens da Base 2 são mostradas na Figura 5.7.

- Base 3: foi obtida da mesma forma que a Base 2. Foram geradas 4666 leituras do laser e 1182 imagens RGB e dados de profundidade, sendo

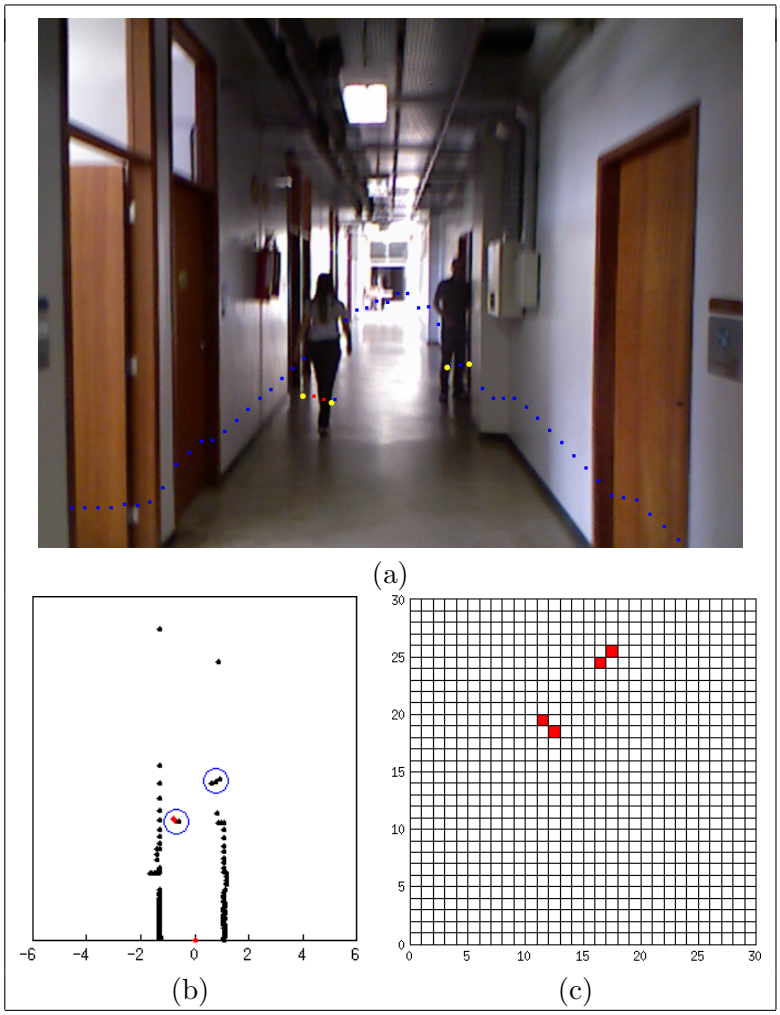


Figura 5.6: Exemplo de mapeamento de pontos do laser e de células da grade para a imagem. (a) Imagem contendo duas pessoas. Os pontos do laser são representados pelos pontos vermelhos (pessoas) e azuis (não pessoas), de acordo com a classificação do detector de Bellotto e Hu [2009]. Os centros das células da grade onde existem pessoas são representados pelos círculos amarelos. (b) A leitura do laser no ambiente mostrado em (a). Os pontos do laser que correspondem a pessoas são destacados pelas circunferências. (c) Grade de ocupação semântica mostrando a presença de pessoas nas células de cor vermelha. As duas pessoas na cena ocupam, cada uma, o espaço de duas células.

Figura 5.7: Exemplos de imagens da Base 2. Nessa base, a plataforma MARIA movia-se pelos corredores do prédio da Escola de Engenharia/UFMG passando por várias pessoas que encontravam-se no ambiente.

que após a subamostragem restaram 394 imagens.

Para permitir uma análise quantitativa dos experimentos, foi realizada a anotação das bases de dados. A anotação consistiu em rotular os dados dos sensores como pessoas e não pessoas, gerando um *ground-truth* das Bases 1 e 2. O *ground-truth*[3] da base de dados indica, tanto para os dados do laser quanto da câmera, a localização das pessoas e é utilizado para a avaliação dos detectores e da metodologia proposta.

A maioria dos trabalhos revisados no Capítulo 3 realiza a anotação manual das imagens ou das leituras do laser (como [Bellotto e Hu, 2009]). Em [Premebida et al., 2009], os dados do laser são manualmente validados e a base de imagens foi automaticamente extraída pelo mapeamento das ROI (regiões de interesse) detectadas pelo laser, estando sujeito aos erros do sistema do laser. Em [Oliveira et al., 2010b] ocorreu o processo inverso, em que as imagens foram anotadas manualmente, gerando as posições onde haviam pessoas. As posições foram projetadas no espaço do laser, que teve suas leituras anotadas. Esse método está sujeito a erros humanos além da inconsistência entre anotadores.

Nesta tese, para a anotação da Base 1, como o fundo era estático, a anotação do laser foi realizada automaticamente, de forma similar a [Mozos et al., 2010], que utiliza uma técnica de subtração do fundo. Para rotular os pontos do laser como positivos ou negativos, considerou-se, no ambiente onde foram realizados os experimentos, um espaço vazio em frente aos sensores. Os pontos foram automaticamente rotulados como positivos se estavam dentro da região vazia e como negativos se estavam fora dessa região, sendo que foram considerados apenas a região dentro do campo de visão da câmera. Esse método possibilita a anotação dos dados do laser de forma mais rápida

[3] *Ground-truth* é também conhecido como gabarito, verdade terreno ou padrão de ouro.

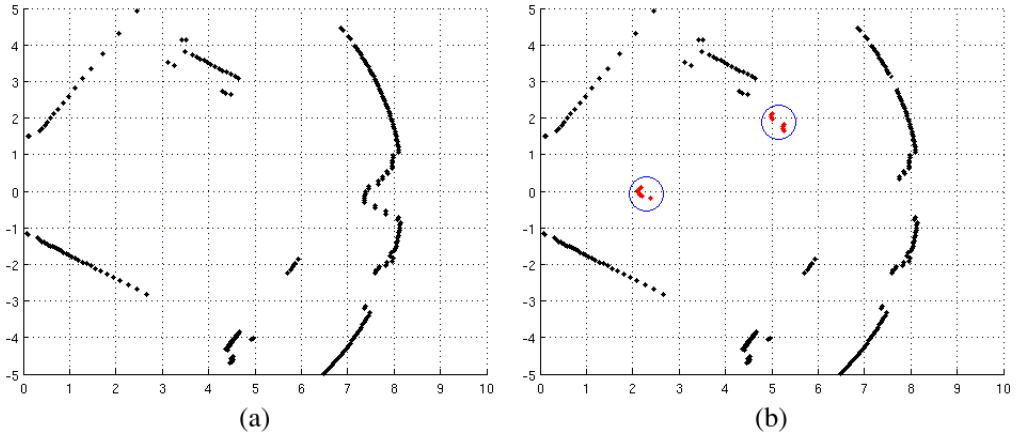


Figura 5.8: Pontos da leitura do laser no ambiente da Base 1, que está posicionado em $(0, 0)$. (a) Leitura do laser no ambiente onde foi produzida a Base 1, sem a presença de pessoas. (b) Leitura do mesmo ambiente, porém com a presença de pessoas, representadas pelos pontos que foram marcados por uma circunferência.

e sem a necessidade de rotular cada ponto manualmente, porém é necessário o conhecimento do ambiente. A Figura 5.8(a) mostra uma leitura do laser no ambiente onde foi produzida a Base 1, sem a presença de pessoas. Essa leitura serviu de referência para rotular os pontos marcados por uma circunferência em (b), que são relativos a pessoas. Como resultado, 5960 pontos foram rotulados como pessoas.

As imagens da câmera da Base 1 foram anotadas manualmente. Os quadros foram anotados com o auxílio de um programa desenvolvido no Matlab, que mostra a visualização de cada imagem e solicita ao usuário que marque com o mouse os vértices dos retângulos que delimitam as pessoas, gravando as coordenadas em arquivo. A anotação das imagens inclui todas as pessoas visíveis, exceto as mais distantes (aproximadamente 20 metros), que são representadas por poucos pixeis na imagem e cuja visualização demanda uma maior atenção. Como resultado foram obtidas 3889 pessoas anotadas na base de imagens.

Para a Base 2, a anotação ocorreu apenas nas imagens. O mesmo método para a anotação das imagens da Base 1 foi utilizado, resultando em 2264 pessoas anotadas. Entretanto, a anotação não foi realizada em todas as imagens da base pois apenas um subconjunto foi utilizado para os experimentos. Desse subconjunto contendo 2746 imagens, foram removidas da anotação as imagens que continham pessoas com os pés oclusos, pois o mapeamento das pessoas do referencial da imagem para o referencial da grade considera explicitamente a posição dos pés das pessoas, como explicado na Seção 4.4, resultando em 2264 pessoas anotadas em 1600 imagens.

5.2 Experimentos

Os detectores de pessoas escolhidos para os experimentos são listados a seguir:

- Detector 1: detector de pernas nos dados do laser chamado *libdetect* de Bellotto e Hu [2009], disponível em `http://webpages.lincoln.ac.uk/nbellotto/software/libdetect_doc/`. O detector foi utilizado com os parâmetros padrões.

- Detector 2: detector de pernas nos dados do laser chamado *People2D* de Spinello e Siegwart [2008], disponível em `http://www2.informatik.uni-freiburg.de/~spinello/people2D.html`. Como o detector requer treinamento, para os experimentos com a Base 1, este foi realizado utilizando-se a base fornecida pelos mesmos autores, porém subamostrada de $0,5°$ para $1°$ de resolução angular. Para os experimentos com a Base 2, o treinamento foi realizado utilizando-se os dados do laser e o *ground-truth* da Base 1.

- Detector 3: detector de pessoas em imagens chamado ACF (*Aggregate Channel Features*) de Dollar et al. [2014], disponível em `http://vision.ucsd.edu/~pdollar/toolbox/doc/`. Nos experimentos seus resultados foram filtrados devido a várias detecções incorretas na parte superior da imagem e outras detecções que englobavam uma área muito grande da imagem, sendo que as detecções acima de 180 pixeis em relação ao eixo y da imagem e maiores que 420 pixeis de altura foram removidas. O detector foi utilizado com o treinamento e parâmetros padrões.

- Detector 4: detector de pessoas em imagens de Dalal e Triggs [2005], disponível em `http://opencv.org/`. O detector foi utilizado com o treinamento e parâmetros padrões.

- Detector 5: detector de pessoas em dados de profundidade do Kinect fornecido pelo pacote do ROS *OpenNi tracker*, que utiliza um algoritmo de rastreamento da posição de articulações do corpo humano (esqueleto). O pacote foi utilizado com seus parâmetros padrões.

5.2.1 Detectores individuais

Os experimentos realizados para estudo dos sensores e algoritmos de detecção de pessoas consistiu em aplicar os Detectores 1, 2, 3 e 4 na Base 1 (dois detectores nos dados do laser e dois detectores nas imagens) e comparar seus resultados com o *ground-truth* da base. Também foram comparados o *ground-truth* dos dados laser com o *ground-truth* das imagens para verificar o comportamento dos sensores em relação à indicação da presença das pessoas que estavam no ambiente. Os experimentos são detalhados a seguir:

1. Comparação do *ground-truth* das leituras do laser e das imagens, por

meio da projeção dos pontos de laser marcados como pessoas nas imagens utilizando as informações da calibração. O objetivo é mostrar que o número de pessoas anotadas nas leituras do laser, mesmo considerando apenas o campo de visão da câmera, difere do número de pessoas anotadas nas imagens. Neste experimento e nos seguintes que foram avaliados em relação à anotação das imagens, foi adicionada uma tolerância de 15 pixeis antes e depois do retângulo demarcado na anotação da imagem devido a erros de sincronização, movimento das pessoas e calibração.

2. Detecção de pernas no laser utilizando o Detector 1 avaliada em relação à anotação do laser e também em relação à anotação das imagens. Toda a base do laser foi utilizada para teste, já que o algoritmo não requer treinamento.

3. Detecção de pernas no laser utilizando o Detector 2 avaliada com base na anotação do laser e também em relação à anotação das imagens. A base do laser foi dividida em duas partes: a primeira, com 719 leituras, foi usada para determinar o limiar da classificação (limiar $= -0,12$) e a segunda parte, com 2000 leituras, foi utilizada para teste.

4. Detecção de pessoas nas imagens utilizando o Detector 3. Como resultado da detecção, são obtidas as coordenadas do retângulo englobando a região da imagem considerada como pessoa pelo classificador. Para comparar com a anotação das imagens, foi calculada a área de interseção entre o retângulo da detecção e o retângulo da anotação, sendo que a razão entre a área de interseção e a área do retângulo da anotação precisa ser maior que um limiar para a detecção ser considerada como verdadeiro positivo. O limiar escolhido foi $0,5$, o mesmo que no

trabalho de Dollar et al. [2012], porém nesse trabalho a razão é entre a área de interseção e a união das áreas do retângulo da detecção e do retângulo da anotação, o que faz com que a avaliação seja mais rigorosa em alguns casos em relação à localização dos retângulos (como o objetivo do experimento é verificar a presença de pessoas, foi tolerado um maior erro na localização).

5. Detecção de pessoas nas imagens utilizando o Detector 4. Para comparar com a anotação das imagens, foi utilizado o mesmo critério explicado para o Detector 3.

As seguintes métricas foram utilizadas para avaliar os resultados dos experimentos:

- Porcentagem de verdadeiros positivos (VP): porcentagem de detecções corretas em relação ao total de pessoas anotadas.
- Porcentagem de falsos negativos (FN): 1 - VP.
- Porcentagem de falsos positivos (FP): porcentagem de detecções incorretas em relação ao total de pessoas detectadas ou (1 - Precisão).
- Precisão: razão entre o número de detecções corretas e o total de detecções, isto é, $VP/(VP + FP)$. A precisão é uma medida do número de acertos em relação ao total de objetos identificados pelo classificador como positivos.
- FPPI: número de falsos positivos por imagem.

O tempo médio de execução, em um computador Intel Core2 Duo $1,8$ GHz, também foi calculado para os detectores de pessoas do laser e da câmera, sendo considerado como tempo inicial a partir do momento em que

o dado está disponível para processamento e como tempo final o momento em que o detector fornece a resposta da classificação. Os resultados dos experimentos são mostrados na Tabela 5.1.

Tabela 5.1: Resultados dos experimentos preliminares.

Métrica	VP	FN	FP	Prec.	FPPI	Tempo médio (s)
Anotação do laser × anotação das imagens	0,68	0,32	0,10	0,90	0,17	-
Detector 1 × **anotação das imagens**	0,39	0,61	0,12	0,88	0,12	-
Detector 2 × **anotação das imagens**	0,49	0,51	0,21	0,79	0,29	-
Detector 3 × **anotação das imagens**	0,26	0,74	0,10	0,90	0,07	0,28
Detector 4 × **anotação das imagens**	0,35	0,65	0,19	0,81	0,19	0,97
Detector 1 × **anotação do laser**	0,39	0,61	0,27	0,73	-	$3,0 \times 10^{-4}$
Detector 2 × **anotação do laser**	0,97	0,03	0,34	0,66	-	$2,4 \times 10^{-4}$

Os resultados mostraram que há diferenças entre as anotações e as detecções do laser e da câmera. Quando a anotação do laser é comparada com a anotação da câmera, nota-se que as marcações de pessoas em uma não condizem completamente com as marcações da outra, ou seja, somente cerca de 68% das pessoas anotadas nas imagens são as mesmas anotadas no laser. Por exemplo quando uma pessoa está atrás de uma caixa de tamanho médio, a câmera consegue visualizar a parte superior do seu corpo, mas um laser posicionado na altura das pernas somente detectará a caixa. Outro exemplo são falhas que podem ser causadas pela baixa refletividade de roupas pretas a grandes distâncias [Oliveira et al., 2010b], reflexões e efeitos da luz direta do sol [Spinello e Siegwart, 2008]. Cabe observar que, além das falhas do laser já citadas, essa diferença deve-se também a pessoas fora do alcance do laser que estavam visíveis na imagem, anotação minuciosa das imagens que incluiu não somente pessoas inteiras mas também partes do corpo quando havia oclusão parcial da pessoa, restrições da implementação em relação às bordas da ima-

(a) (b) (c)

Figura 5.9: Resultados das detecções. (a) Detecção baseada em visão. (b) Detecção do laser projetada na imagem (círculos) e anotação da imagem em azul. (c) Detecções do laser destacadas pelas circunferências. Nota-se que a pessoa da esquerda vista nas imagens não é percebida pelo laser.

gem (uma pessoa parcialmente fora do campo de visão da câmera pode gerar um ponto médio do laser fora dos limites da imagem, não sendo computada) e segmentação imperfeita da anotação do laser para projeção na imagem.

Como esperado, os resultados das detecções do laser, quando avaliados com a anotação das imagens também tiveram seus resultados prejudicados. Nos resultados de detecção no laser comparados com a anotação das imagens há um número considerável de falsos negativos (FN), que podem ser causados também por pessoas visíveis na câmera mas oclusas no laser, como mostrado na Figura 5.9.

Analisando-se a precisão dos resultados, observa-se que, apesar do Detector 2 ser menos preciso que o Detector 1, na anotação dos dados do laser cujo FOV é mais amplo que o das imagens, ele detecta mais de 96% das pessoas presentes na cena. Nos resultados do Detector 2 $\times$ anotação do laser), observou-se que este detector tende a classificar um maior número de amostras como pessoas ao custo de uma alta taxa de falsos positivos e consequentemente menor precisão. O Detector 1 foi melhor no FOV da imagem em relação à precisão, o que indica que nessa região sua detecção é mais precisa embora o número de VP seja similar ao da anotação com dados do laser.

Em relação a HOG × anotação das imagens, esse detector mostrou-se mais seletivo recuperando menos pessoas porém com menos falsos positivos. Os valores de FPPI dos detectores avaliados em relação à anotação das imagens mostraram valores satisfatórios em relação ao número médio de pessoas por imagem, que é $2,3$, ou seja, os detectores deixam de detectar poucas pessoas presentes na cena.

Os detectores do laser, quando comparados com a anotação do laser, mostraram bons resultados com baixo tempo de execução. Já os detectores baseados em visão, apesar de apresentarem precisão acima de $0,80$, tiveram muitos falsos negativos. Neste caso, o treinamento padrão utilizado pode ter influenciado os resultados pois as bases de imagens originais não possuem as mesmas configurações do experimento, tais como tamanho das pessoas nas imagens. Entretanto, mesmo com um melhor treinamento, o tempo de execução da detecção em imagem é alto em comparação à detecção no laser.

Esses experimentos preliminares mostraram que os resultados de detectores de laser e câmera sozinhos podem não refletir a situação real. Apesar dos bons resultados da detecção do laser em relação à anotação do laser, este detector deixa de detectar pessoas existentes na cena e que a câmera teria condições de detectar (e vice-versa). A escolha dos melhores detectores depende da aplicação, pois para cada uma deve ser avaliado o compromisso entre VP, FN e FP. Por exemplo, em uma aplicação relacionada a busca e resgate de pessoas, uma alta taxa de VP é desejável, mesmo havendo falsos alarmes, pois é importante nesta situação detectar o máximo de pessoas possível. Por outro lado, em uma aplicação de sistemas de apoio ao motorista, uma taxa de VP mais baixa pode não ser tão crítica desde que a taxa de FP não seja alta, evitando um excesso de falsos alarmes que podem perturbar o motorista.

Um dos principais objetivos das abordagens de fusão que detectam candidatos a pessoas com o laser e realizam a classificação das ROIs mapeadas com algoritmos de visão é a redução no tempo de processamento, pois o processamento da imagem é custoso em relação ao tempo e o espaço de busca do algoritmo de visão é reduzido com a fusão. Entretanto, os experimentos desta tese mostraram que a detecção com o laser e com a câmera estão sujeitas a falhas, comprometendo o resultado final. Dessa forma, conclui-se que a solução mais adequada para aproveitar as informações de ambos os sensores é combinar o resultado das detecções em paralelo (Figura 3.4(a) e (b)) e não da forma sequencial (Figura 3.4(c)), em termos de maximizar os resultados da detecção de pessoas. Os resultados dos experimentos (com a exceção do Detector 3, que foi testado posteriormente) foram publicados no trabalho [Batista e Pereira, 2013].

5.2.2 Fusão de detectores

Foram realizados três experimentos com a metodologia proposta para validar todas etapas da metodologia qualitativamente e quantitativamente. Cada experimento foi realizado com uma das bases. A implementação descrita na Seção 4.4 foi utilizada com combinações distintas dos Detectores 1, 2, 3, 4 e 5, construindo como resultado uma grade de ocupação semântica que representa o espaço de trabalho local de um robô de serviço. A Figura 5.10 ilustra esta representação com a plotagem dos centros das células da grade em uma das imagens da base.

No experimento com a Base 1, foram utilizados os Detectores 1, 2 e 4, com os sensores estáticos. Os parâmetros utilizados, descritos nas equações do Capítulo 4, são mostrados na Tabela 5.2. Esses parâmetros foram os que apresentaram melhor compromisso de acordo com o treinamento explicado

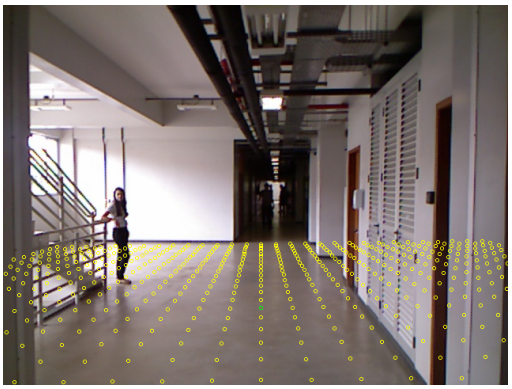


Figura 5.10: Imagem da Base 2 com os centros das células da grade representados pelas circunferências.

na Subseção 5.2.4.

No experimento com a Base 2, foram utilizados os Detectores 1, 2 e 3, em diversas combinações distintas. O Detector 4, embora tenha apresentado uma taxa de verdadeiros positivos maior que o Detector 3 no experimento com a Base 1, não foi utilizado por não apresentar resultados satisfatórios em estudo comparativo de Dollar et al. [2012], além de possuir um tempo de execução maior. Os parâmetros utilizados são mostrados na Tabela 5.3, obtidos com o treinamento explicado na Subseção 5.2.4.

No experimento com a Base 3, os Detectores 1, 3 e 5 foram utilizados. Os parâmetros utilizados são mostrados na Tabela 5.4 e foram obtidos do treinamento do Experimento 2 e, para o Detector 5, foram utilizados os valores de precisão e o valor médio de revocação encontrados no trabalho de Shotton et al. [2011].

A seguir são apresentados resultados qualitativos dos experimentos.

5.2.3 Análise qualitativa

Para ilustrar alguns resultados da metodologia proposta de forma qualitativa, são apresentados exemplos da Base 2 e da Base 3, por possibilitarem

Tabela 5.2: Parâmetros utilizados no experimento com a Base 1.

Parâmetro	**Valor**
μ	$1,46\,\mathrm{m/s}$
σ	$0,63\,\mathrm{m/s}$
$p_{\text{imóvel}}$	$0,5$
Δt	$0,1\,\mathrm{s}$
VPP (Detector 1)	$0,70$
TFN (Detector 1)	$0,10$
VPP (Detector 2)	$0,90$
TFN (Detector 2)	$0,10$
VPP (Detector 4)	$0,70$
TFN (Detector 4)	$0,50$

Tabela 5.3: Parâmetros utilizados nos experimentos com a Base 2. As combinações dos detectores foram: Detectores 1, 2 e 3 (D1D2D3), Detectores 1 e 2 (D1D2), Detectores 1 e 3 (D1D3) e Detectores 2 e 3 (D2D3).

Parâmetro	**D1D2D3**	**D1D2**	**D1D3**	**D2D3**
μ	$1,46\,\mathrm{m/s}$	$1,46\,\mathrm{m/s}$	$1,46\,\mathrm{m/s}$	$1,46\,\mathrm{m/s}$
σ	$0,63\,\mathrm{m/s}$	$0,63\,\mathrm{m/s}$	$0,63\,\mathrm{m/s}$	$0,63\,\mathrm{m/s}$
$p_{\text{imóvel}}$	$0,85$	$0,85$	$0,85$	$0,85$
Δt	$0,1\,\mathrm{s}$	$0,1\,\mathrm{s}$	$0,1\,\mathrm{s}$	$0,1\,\mathrm{s}$
VPP (Detector 1)	$0,95$	$0,95$	$0,95$	-
TFN (Detector 1)	$0,30$	$0,50$	$0,30$	-
VPP (Detector 2)	$0,50$	$0,40$	-	$0,50$
TFN (Detector 2)	$0,40$	$0,50$	-	$0,40$
VPP (Detector 3)	$0,95$	-	$0,95$	$0,95$
TFN (Detector 3)	$0,50$	-	$0,50$	$0,50$

Tabela 5.4: Parâmetros utilizados no experimento com a Base 3.

Parâmetro	**Valor**
μ	$1,46\,\mathrm{m/s}$
σ	$0,63\,\mathrm{m/s}$
$p_{\text{imóvel}}$	$0,5$
Δt	$0,1\,\mathrm{s}$
VPP (Detector 1)	$0,95$
TFN (Detector 1)	$0,30$
VPP (Detector 3)	$0,95$
TFN (Detector 3)	$0,50$
VPP (Detector 5)	$0,75$
TFN (Detector 5)	$0,50$

experimentos com todas as etapas da metodologia devido ao movimento do robô.

A Figura 5.11 mostra a execução de dois instantes consecutivos do experimento com a Base 2. O primeiro instante é representado pelas figuras (a) e (c)-(j) e o segundo instante pelas figuras (b) e (k)-(r). As figuras 5.11(a) e (b) mostram imagens da câmera com as pessoas detectadas marcadas por retângulos. A grade nas figuras (c) e (k) representam o *ground-truth*, que foi anotado manualmente utilizando os dados não processados do laser projetados na grade. Nessa grade, apenas pessoas foram anotadas e são representadas pelas células vermelhas. Como mostrado nas figuras (c) e (k), existem duas pessoas na cena: uma posicionada acima e à esquerda (Pessoa 1), que ocupa quatro células nas grades da anotação; e outra pessoa abaixo e à direita (Pessoa 2), que ocupa duas células na figura (c) e três na figura (k). A Pessoa 1 e a Pessoa 2 estão se movendo em direções opostas. Como o campo de visão da câmera é menor que o do laser, as imagens (a) e (b) não mostram a Pessoa 2. Observa-se que, nas figuras (a) e (b), existem outras pessoas detectadas pelo Detector 3 que não foram representadas na grade por estarem a mais de 9

metros de distância do robô, e portanto, fora dos limites da grade.

As figuras 5.11(d) e (l) mostram a etapa de predição baseada no movimento das pessoas. Nessas figuras, quanto mais escura for a célula, maior é a probabilidade de estar ocupada por pessoas. Como a etapa de predição é baseada na crença calculada no intervalo de tempo anterior (probabilidade *a priori*), (l) é calculada pela aplicação do modelo de movimento das pessoas à grade da Figura 5.11(j), que é o resultado final da metodologia proposta na primeira iteração mostrada na Figura 5.11. A probabilidade *a priori* relacionada à Figura 5.11(d) não é mostrada. Comparando-se (j) e (l), nota-se que o modelo de movimento das pessoas suaviza a probabilidade em todas as células da grade. A etapa de predição considera que a probabilidade de existência de pessoas fora da grade é igual a $0,5$, ou seja, não há conhecimento sobre a presença de pessoas nesta região. Dessa forma, o modelo de movimento considera que pessoas fora dos limites da grade podem mover-se para seu interior.

Nas figuras 5.11(e) e (m) são mostrados os resultados da predição baseada no movimento dos sensores. Nestas grades é quase imperceptível o efeito de um pequeno deslocamento do robô, que movia-se a uma velocidade linear de $0,35\,\mathrm{m/s}$ e angular de $0\,\mathrm{rad/s}$, num intervalo de tempo equivalente a $0,11\,\mathrm{s}$.

As grades nas figuras 5.11(f) a (h) e (n) a (p) mostram os resultados obtidos pelos detectores mapeados nas grades semânticas. Nessas grades, as células vermelhas correspondem a pessoas, as células cinzas indicam regiões fora do alcance dos sensores e as células brancas representam espaço livre de pessoas. A Figura 5.11(f) refere-se ao Detector 1 que, utilizando os dados do laser, detectou apenas a Pessoa 2 no primeiro instante de tempo. Na próxima iteração, esse detector encontrou apenas a Pessoa 1 (Figura 5.11(n)). O outro detector baseado em laser, o Detector 2, na Figura 5.11(g) encontrou

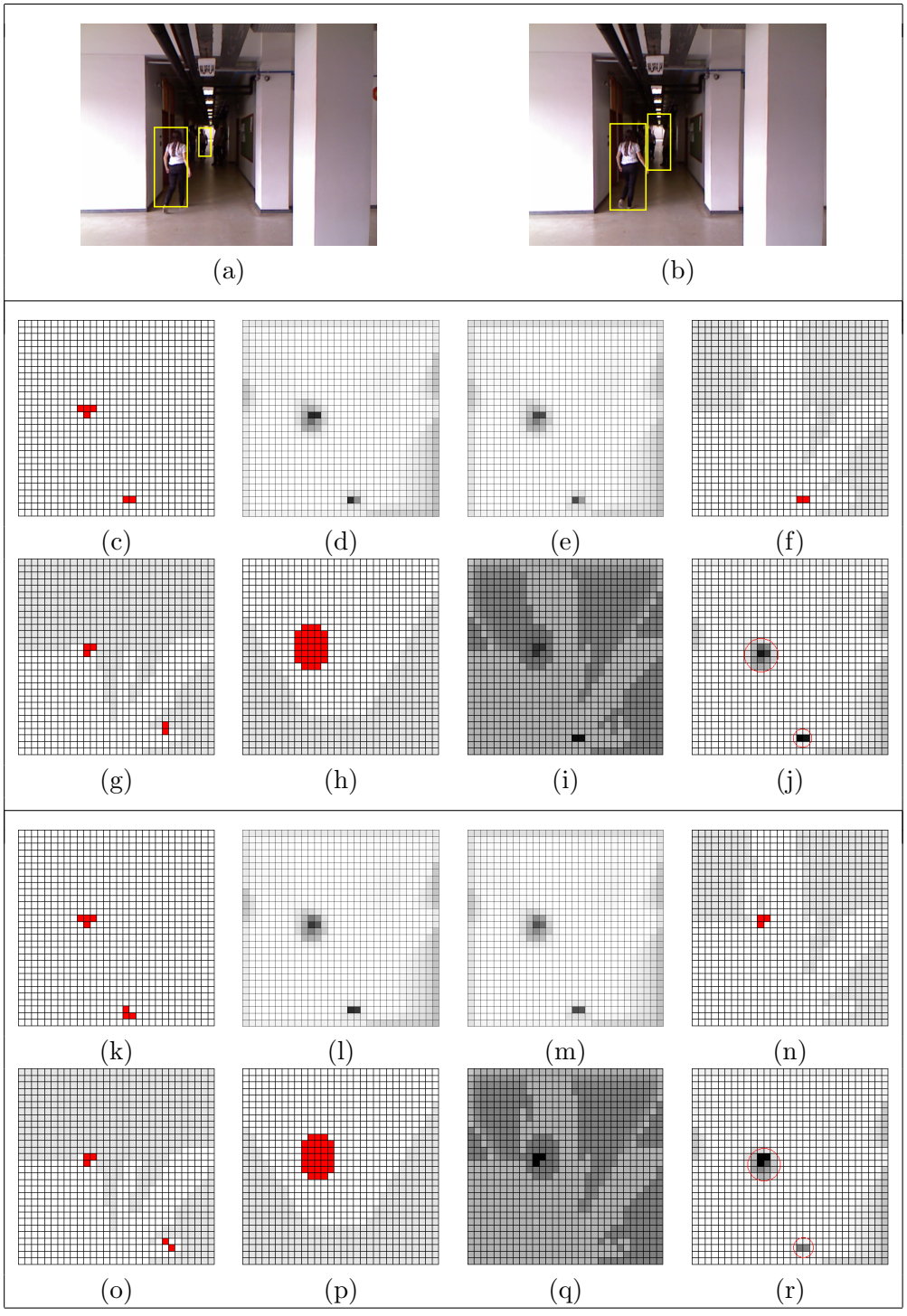


Figura 5.11: Resultados experimentais em dois instantes de tempo: Tempo t ((a) e (c)–(j)); tempo $t+1$ ((b) e (k)–(r)). As figuras mostram: imagens da câmera com as detecções ((a) e (b)); o *ground-truth* com as células contendo pessoas marcadas em vermelho ((c) e (k)); a grade resultante da etapa de predição baseada no movimento das pessoas ((d) e (l)); a grade resultante da predição baseada no movimento dos sensores ((e) e (m)); as detecções do Detector 1 usando laser ((f) e (n)); as detecções do Detector 2 usando laser ((g) e (o)); as detecções do Detector 3 usando imagens ((h) e (p)); a fusão dos detectores ((i) e (q)); e a grade final ((j) e (r)).

a Pessoa 1 mas também detectou de forma incorreta outra pessoa na parte inferior direita da grade. No próximo instante de tempo, esse detector novamente detectou a Pessoa 1 e uma outra pessoa, que não foi uma detecção correta (Figura 5.11(o)). As figuras 5.11(h) e (p) mostram os resultados do Detector 3, que é baseado em imagens. Este detector encontrou com sucesso a Pessoa 1. A Pessoa 2 não poderia ser detectada usando imagens, pois está fora do campo de visão da câmera. Nas figuras (h) e (p), a detecção é representada por várias células devido à grande incerteza em relação à posição das pessoas detectadas com apenas uma câmera.

A fusão dos três detectores usando as equações (4.23) e (4.24) é mostrada nas figuras 5.11(i) e (q). Novamente, as células escuras indicam uma maior probabilidade de existir pessoas. As células livres possuem uma probabilidade pequena, que é representada por cinza claro. No primeiro intervalo de tempo, a Pessoa 1 e a Pessoa 2 foram detectadas pelo menos por um dos detectores e portanto a fusão mostra uma alta probabilidade nas células ocupadas pelas pessoas (Figura 5.11(i)). Por outro lado, nos resultados da fusão do segundo intervalo de tempo mostrado na Figura 5.11(q), apenas a posição ocupada pela Pessoa 1 possui uma probabilidade alta, pois a Pessoa 2 não foi detectada neste instante de tempo. Uma vantagem mostrada pela metodologia proposta é que, graças à baixa confiança atribuída ao Detector 2, os falsos positivos (isto é, detecções incorretas) tiveram suas probabilidades minimizadas durante a etapa de fusão em ambos instantes de tempo.

A probabilidade *a posteriori*, calculada usando a Equação (4.22), que combina a crença obtida na etapa de predição com a probabilidade obtida da fusão dos detectores, é mostrada nas figuras 5.11(j) e (r). Na Figura 5.11(j), as células ocupadas por pessoas possuem as maiores probabilidades, que representam a localização mais provável das pessoas. As figuras mostram que,

quando dois detectores de pessoas diferentes detectam a mesma pessoa, a probabilidade das células relacionadas a esta detecção é aumentada.

A Figura 5.11(r) é o resultado da grade após a etapa de atualização no segundo instante de tempo. A localização da Pessoa 1 (no topo à esquerda) é aprimorada pelo aumento da probabilidade nas células em que os detectores baseados em laser encontraram essa pessoa. Embora os três detectores falharam em detectar a Pessoa 2, as células próximas à sua localização real possuem uma probabilidade proeminente (porém menor que na Figura 5.11(j)) devido ao fato de que, na etapa de atualização da metodologia, a informação da etapa de predição é também considerada, pois é baseada na crença calculada no intervalo de tempo anterior.

Além de possibilitar uma filtragem nos dados de entrada, a metodologia proposta tem a vantagem de permitir a adição ou remoção de detectores e por sua vez sensores de forma facilitada. A metodologia proposta não limita o número de detectores permitidos, sendo que a adição de mais um detector, que pode ou não utilizar dados de um novo sensor, implica em:

- Obter o modelo do detector: deve ser obtido o erro de localização das detecções e os parâmetros VPP e TFN para o detector mais adequados para a combinação com os outros detectores do sistema. Para isso é necessário treinamento com dados obtidos em que a localização das pessoas é conhecida.

- Fornecer para o detector dados do sensor específico para o detector, que deverá retornar a posição das pessoas no referencial do sensor.

- Transformar coordenadas das pessoas para o referencial da grade, projetando as pessoas nas células correspondentes conforme o modelo de erro.

- Combinar a grade com as detecções realizadas pelo novo detector com as grades das detecções dos outros detectores do sistema na etapa de atualização.

Quando a adição de mais um detector ocorre em função de um novo sensor no sistema, é necessária realizar a calibração do sensor para encontrar a transformação do referencial do novo sensor para o referencial da grade semântica e sincronizar os dados provenientes do sensor utilizado com dados de outros sensores do sistema. Para ilustrar a escalabilidade da metodologia em relação ao número de detectores utilizados, foi realizado um experimento com a Base 3, que utiliza dados de um sensor a laser, de uma câmera RGB e de profundidade do Kinect. Para cada sensor foi utilizado um detector de pessoas distinto, sendo o Detector 1 para os dados do sensor a laser, o Detector 3 para as imagens da câmera RGB e o Detector 5 para os dados de profundidade do Kinect. O Detector 5 disponibiliza as pessoas detectadas por meio de um esqueleto sob a forma de coordenadas das partes do corpo em relação ao referencial do Kinect. No experimento com a Base 3 foram utilizadas as coordenadas dos pés direito e esquerdo para representar a posição das pessoas. Essas coordenadas foram transformadas para o referencial da grade semântica de acordo com as informações obtidas da calibração. A Figura 5.12 mostra a execução de dois instantes consecutivos do experimento. O primeiro instante é representado pelas figuras (a) e (c)-(j) e o segundo instante pelas figuras (b) e (k)-(r), de forma similar à Figura 5.11. As figuras 5.12(a) e (b) mostram imagens da câmera com as pessoas detectadas marcadas por retângulos. A grade nas figuras (c) e (k) representam o *ground-truth*.

Como mostrado nas figuras (c) e (k), existem duas pessoas na cena: uma posicionada acima e à esquerda (Pessoa 1) e outra pessoa abaixo e à direita (Pessoa 2). A Pessoa 1 e a Pessoa 2 estão se movendo na mesma direção e

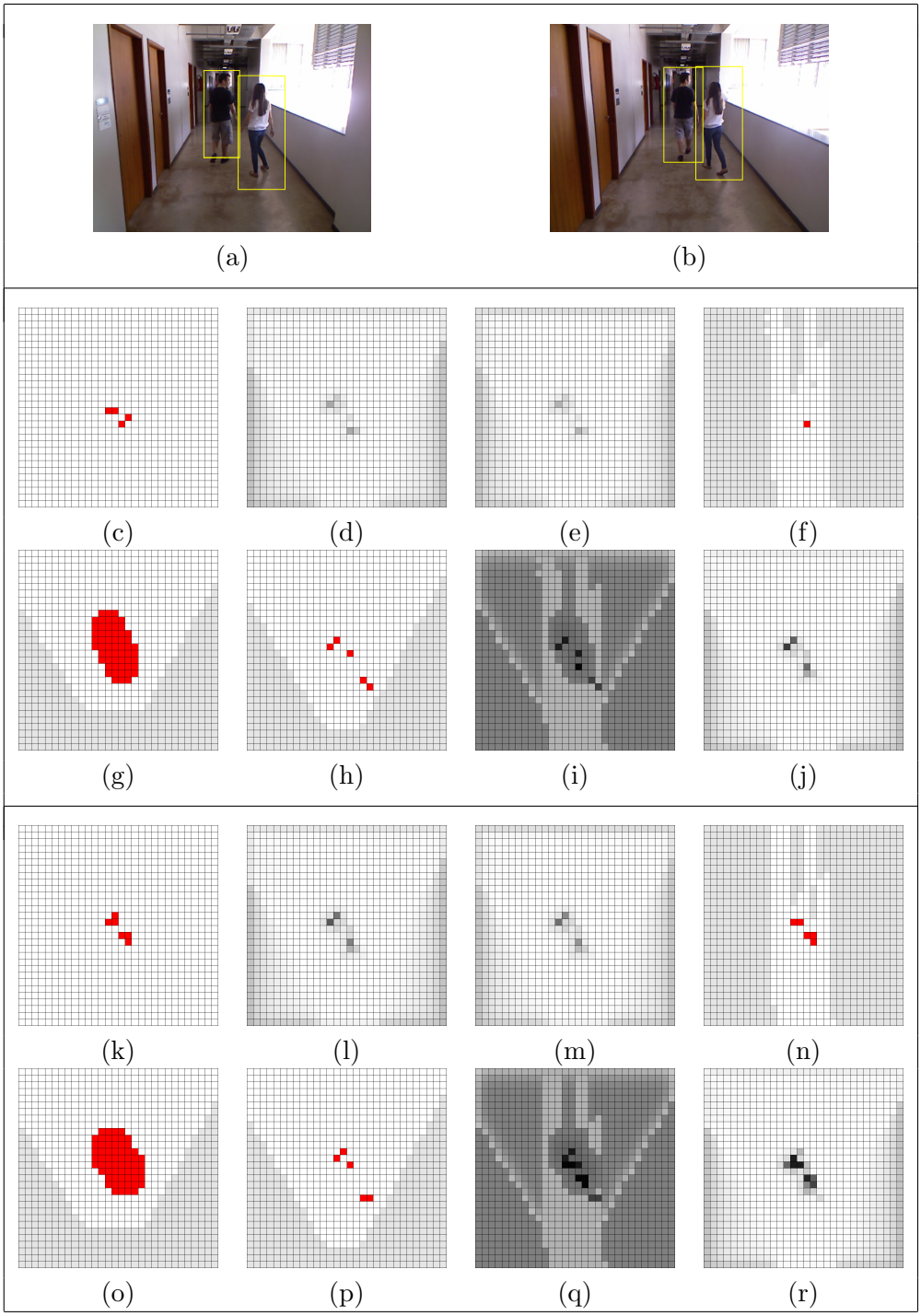


Figura 5.12: Resultados experimentais em dois instantes de tempo: Tempo t ((a) e (c)–(j)); tempo $t+1$ ((b) e (k)–(r)). As figuras mostram: imagens da câmera RGB com as detecções ((a) e (b)); o *ground-truth* com as células contendo pessoas marcadas em vermelho ((c) e (k)); a grade resultante da etapa de predição baseada no movimento das pessoas ((d) e (l)); a grade resultante da predição baseada no movimento dos sensores ((e) e (m)); as detecções do Detector 1 ((f) e (n)); as detecções do Detector 3 ((g) e (o)); as detecções do Detector 5 ((h) e (p)); a fusão dos detectores ((i) e (q)); e a grade final ((j) e (r)).

estão de costas para os sensores. As figuras 5.12(d) e (l) mostram a etapa de predição baseada no movimento das pessoas e nas figuras 5.12(e) e (m) são mostrados os resultados da predição baseada no movimento dos sensores. As grades nas figuras 5.12(f) a (h) e (n) a (p) mostram os resultados obtidos pelos detectores mapeados nas grades semânticas. Como no exemplo anterior, nessas grades, as células vermelhas correspondem a pessoas, as células cinzas indicam regiões fora do alcance dos sensores e as células brancas representam espaço livre de pessoas.

A Figura 5.12(f) refere-se ao Detector 1 que, utilizando os dados do laser, detectou apenas a Pessoa 1 no primeiro instante de tempo. Na próxima iteração, esse detector encontrou as duas pessoas (Figura 5.12(n)). O detector baseado em imagens da câmera RGB, o Detector 3, nas Figura 5.12(g) e (o) encontrou as duas pessoas em ambos os instantes de tempo. As detecções foram unidas na grade pois, além de estarem próximas, cada detecção é representada por várias células devido à grande incerteza em relação à posição das pessoas detectadas. As figuras 5.12(h) e (p) mostram os resultados do Detector 5, que é baseado nos dados de profundidade do Kinect. Este detector encontrou três pessoas em cada instante de tempo. Nas figuras (h) e (p), uma das detecções é um falso positivo e as outras detecções estão localizadas próximas ou na mesma posição das pessoas do *ground-truth*. A fusão dos três detectores é mostrada nas figuras 5.12(i) e (q). A probabilidade *a posteriori* é mostrada nas figuras 5.12(j) e (r). Pode-se observar que as células ocupadas de fato por pessoas possuem probabilidades destacadas e os falsos positivos encontrados pelo Detector 5 foram filtrados pela metodologia.

Há alguns fatores que podem impactar negativamente nos resultados da metodologia proposta, tais como a densidade de pessoas na cena e consequentemente a oclusão de partes do corpo, pessoas com movimento muito atípico,

a frequência com que os detectores enviam informações sobre o ambiente, a imprecisão dos detectores na localização de pessoas e detectores que deixam de encontrar um grande número de pessoas ou que detectam muitos falsos positivos. Para ilustrar uma situação em que a grade resultante da etapa de atualização não corresponde à situação real, a Figura 5.13 traz uma iteração da Base 2 em que a detecção incorreta do Detector 1, que possui uma uma confiança alta, faz com que a probabilidade da célula correspondente à localização do falso positivo seja aumentada. Essa figura mostra os resultados dos três detectores mapeados para a grade, a imagem da cena (que contém apenas uma pessoa) com o resultado do Detector 3 marcado com um retângulo e a grade de atualização resultante. Nota-se, na Figura 5.13(e), que as probabilidades das células relacionadas às detecções estão destacadas em relação à vizinhança. A probabilidade da detecção correta é reforçada pois os três detectores encontram a pessoa na cena, entretanto, o falso positivo detectado pelo Detector 1, que possui uma confiança alta, também resulta em uma probabilidade destacada na grade.

O resultado final da metodologia proposta é a probabilidade *a posteriori* e, na implementação utilizada neste experimento, consiste de uma grade em que cada célula representa uma região do espaço e seus valores representam a probabilidade da célula estar ocupada por pessoas. Para que essa informação seja útil em aplicações reais, é necessário que a grade final passe pela etapa de classificação para determinar se pessoas foram realmente detectadas com base nas probabilidades resultantes e obter sua localização. Para possibilitar uma análise quantitativa dos resultados, foi aplicada a classificação de *blobs* na grade de probabilidades *a posteriori*, explicada na Subseção 4.4.2. Um exemplo desse algoritmo aplicado à grade de probabilidades *a posteriori* da Figura 5.11(r), gerada como resultado da metodologia proposta, foi mostrado

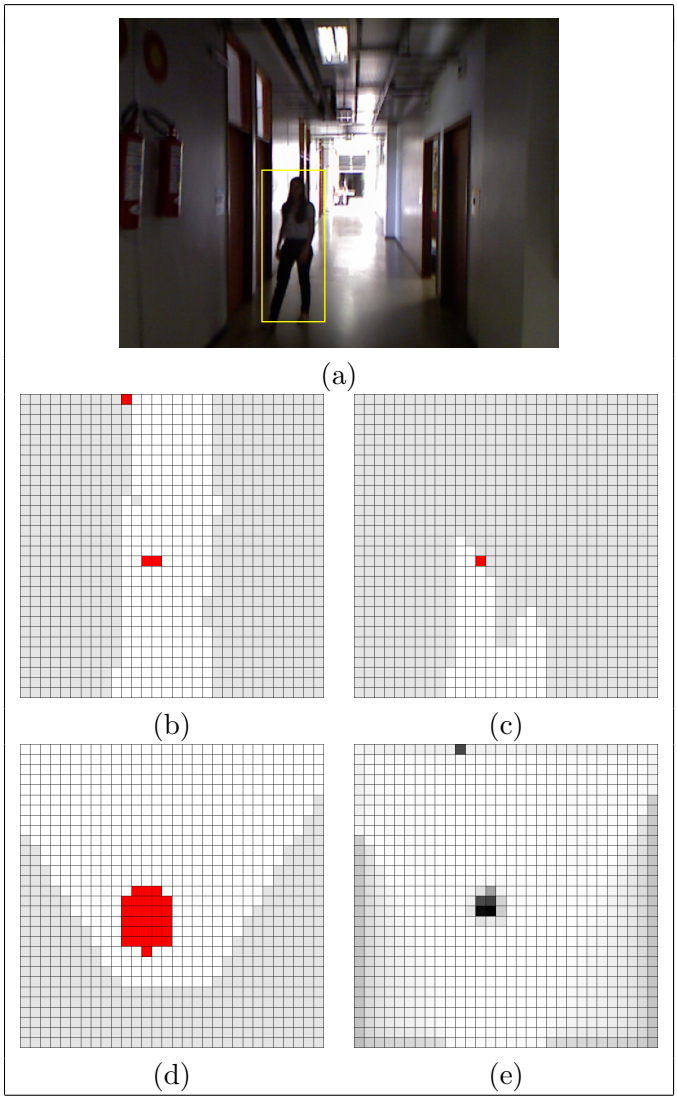


Figura 5.13: Efeito da detecção de um falso positivo no resultado da etapa de atualização. (a) Imagem da Base 2 contendo uma pessoa. O retângulo marca a detecção obtida pelo Detector 3. (b) Resultados do Detector 1 mapeados para a grade. As células correspondentes a pessoas detectadas são marcadas de vermelho, porém a célula vermelha na parte superior da grade corresponde a um falso positivo, pois havia somente uma pessoa na cena. (c) Resultados do Detector 2 mapeados para a grade (células vermelhas). (d) Resultados do Detector 3 mapeados para a grade (células vermelhas). (e) Grade de probabilidades *a posteriori* resultante da etapa de atualização.

nas Figuras 4.8(c).

Uma avaliação qualitativa da metodologia proposta seguida da etapa de classificação por meio de vídeos está disponível na página: `http://coro.cpdee.ufmg.br/movies/peopledetection`. A seguir, será apresentada uma análise objetiva dos seus resultados.

5.2.4 Análise quantitativa

Para possibilitar uma avaliação quantitativa da metodologia proposta, as etapas da abordagem foram executadas seguidas pela classificação baseada em *blobs* na Base 1, sem considerar a etapa de predição baseada no movimento dos sensores e na Base 2, aplicando todas as etapas da metodologia.

No experimento com a Base 1, os dados da base foram divididos entre treino e teste de acordo com a anotação disponível, correspondendo a 600 imagens para treino e 400 para teste. O treinamento foi realizado para obter os melhores parâmetros para serem utilizados pela metodologia nos testes e foi feito apenas com os dados de treino utilizando diversos valores de VVP, TFN, velocidade média, probabilidade da pessoa permanecer parada, dentre outros valores de parâmetros necessários para a metodologia. Essas combinações dos valores de parâmetros foram aplicados nos dados de treino, sendo que o melhor resultado (com melhor revocação mantendo a precisão alta) foi escolhido para novo treinamento dos parâmetros da classificação baseada em *blobs*, variando-se o limiar de agrupamento e o limiar de detecção. Os melhores valores obtidos (limiar de detecção $= 0,75$, limiar de agrupamento $= 0,65$ e dos outros parâmetros são mostrados na Tabela 5.2) foram escolhidos para teste com os dados de teste da Base 1.

Os resultados dos dois detectores baseados em laser, do detector baseado em imagens e da classificação de *blobs* usando a metodologia proposta

Tabela 5.5: Resultados da detecção de pessoas obtidos com os dados de teste da Base 1. Os detectores comparados são: Detector 1 [Bellotto e Hu, 2009], Detector 2 [Spinello e Siegwart, 2008], Detector 4 [Dalal e Triggs, 2005] e metodologia proposta seguida da classificação.

Detector	Revocação	Precisão	FPCPI	Medida-$F1$
Detector 1	$0,47$	$0,58$	$3,03$	$0,52$
Detector 2	$0,56$	$0,54$	$4,82$	$0,55$
Detector 4	$0,09$	$0,37$	$13,37$	$0,14$
Metodologia Proposta com a classificação	$0,56$	$0,55$	$6,65$	$0,56$

foram automaticamente comparados com o *ground-truth* dos dados do laser projetado na grade. A Tabela 5.5 mostra os resultados numéricos usando as seguintes métricas:

- Revocação (*recall*): é a porcentagem de detecções corretas em relação ao número real de pessoas.
- Precisão: já definida anteriormente como a razão entre o número de detecções corretas e o número total de detecções.
- Falsos Positivos em relação às Células Por Imagem (FPCPI): mede a precisão na localização das detecções por meio do número médio de células que o detector considerou ocupadas por pessoas mas que estavam livres no *ground-truth* (quanto menor for o FPCPI, melhor é o resultado).
- Medida-$F1$: média harmônica da precisão e da revocação, dada por $\text{F1} = 2 \times \text{Precisão} \times \text{Revocação}/(\text{Precisão} + \text{Revocação})$, em que resultados mais próximos de 1 são melhores do que mais próximos de zero.

Os resultados para a Base 1 mostraram que a combinação dos resultados

dos detectores usando a metodologia aproveita as vantagens de cada detector gerando um resultado em geral melhor do que os detectores individuais, visto que a medida-$F1$ foi superior e as outras métricas se mantiveram entre os valores obtidos pelos detectores.

No experimento com a Base 2, os dados da base foram divididos entre dados de treino e teste de acordo com a anotação das imagens:

- Treino: dados que correspondem às primeiras 1100 imagens anotadas, com 1885 pessoas.

- Teste: dados que correspondem às últimas 500 imagens anotadas, com 379 pessoas.

Para obter os melhores parâmetros a serem utilizados no experimento com os dados de teste, o treinamento foi realizado com base nos valores de precisão e revocação dos detectores individuais no primeiro experimento com a Base 1, isto é, a partir dos parâmetros obtidos anteriormente foram determinados diversos valores de VVP e TFN para serem testados com a parte de treino da Base 2. Da mesma forma, foram escolhidos valores para os outros parâmetros como velocidade média, probabilidade da pessoa permanecer parada, etc. As combinações dos valores dos parâmetros determinados foram aplicados nos dados de treino, sendo que o melhor resultado (com melhor revocação mantendo a precisão alta) foi escolhido para novo treinamento da classificação baseada em *blobs*, variando-se o limiar de agrupamento e o limiar de detecção. Os melhores valores obtidos foram escolhidos para serem testados com a parte de teste da Base 2. Os resultados do treinamento para os parâmetros Valor Preditivo Positivo (VPP) e Taxa de Falsos Negativos (TFN) são mostrados nos gráficos da Figura 5.14. Os valores do VPP e TFN que apresentam melhor compromisso estão próximos da região em que a revocação é igual a

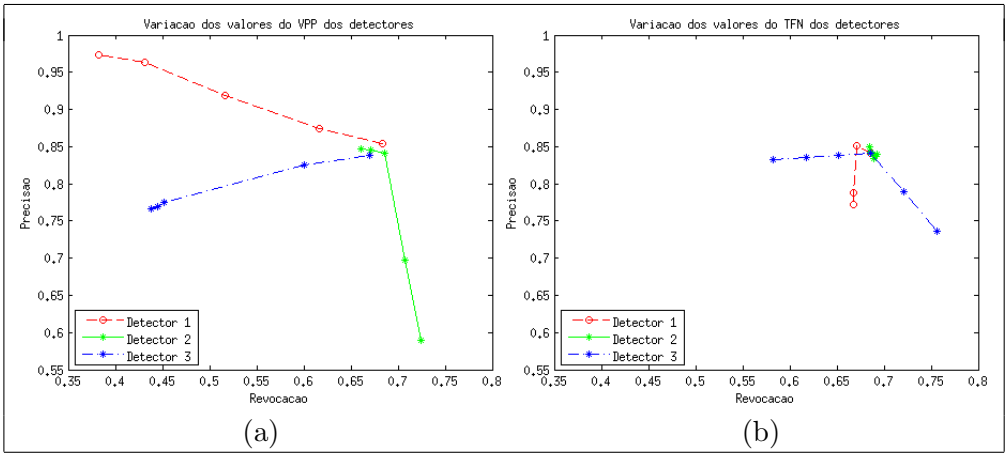


Figura 5.14: Curvas Precisão × Revocação obtidas a partir da variação dos parâmetros VPP e TFN dos detectores com os dados de treino da Base 2. (a) Variação do VPP. (b) Variação do TFN.

$0,7$ e a precisão igual a $0,85$. Esses parâmetros refletem a confiança que se deseja a atribuir aos detectores, por exemplo, um detector cujas informações de localização é bastante precisa pode ter sua confiança aumentada para refletir essa precisão na localização, enquanto que para outro detector com boa precisão de localização mas com resultados de detecção pobres pode ter sua confiança diminuída, sendo o VPP e a TFN pontos de partida para esses ajustes de forma a obter melhores resultados.

Os parâmetros obtidos no treinamento foram utilizados no subconjunto da Base 2 que corresponde ao teste. Esse subconjunto permite uma análise de diferentes situações presenciadas pelo robô, que movia-se a velocidades variáveis. Os resultados da classificação foram comparados com o *ground-truth* das imagens da base. Como explicado na Subseção 5.1.2, foram retiradas da avaliação as iterações em que as imagens continham pés oclusos para tornar a localização das pessoas mais precisa. Entretanto, esta não é uma restrição da metodologia, que permite a utilização de sensores que não fornecem a localização de forma precisa, refletindo esta imprecisão no mapeamento das

pessoas detectadas para a grade. Um exemplo desta situação é mostrado na Figura 5.15. A pessoa detectada na imagem é mapeada para a grade com uma grande incerteza em relação ao eixo y, pois considera-se que a pessoa pode estar em qualquer posição dentro da faixa marcada pelas linhas verdes na Figura 5.15(a). Utilizando este mapeamento na metodologia proposta, observa-se que, embora a localização não seja precisa, a presença da pessoa causa um aumento da probabilidade de uma faixa de células. Em uma aplicação como a navegação de robôs, as células com probabilidade mais alta indicariam regiões em que o robô não poderia passar de forma segura, sob o risco de colidir com uma pessoa. A localização das pessoas em imagens pode ser determinada de forma mais precisa utilizando-se visão estéreo ou mapeando-se as pessoas encontradas nas imagens para os dados não processados do laser, que fornece a distância dos pontos. Porém, a precisão da localização com a visão estéreo se deteriora com o aumento da distância [Sabbagh et al., 2010] e, no caso do laser, as pessoas das imagens podem não coincidir com pontos do laser, conforme discutido na Subseção 5.2.1.

Apesar do *ground-truth* da Base 2 não conter as imagens com as pessoas cujos pés estão oclusos, as imagens com oclusões de outras partes do corpo foram mantidas. Os resultados dos dois detectores baseados em laser, do detector baseado em imagens e da classificação de *blobs* usando a metodologia proposta com quatro combinações distintas de detectores foram automaticamente comparados com o *ground-truth* das imagens projetado na grade, utilizando-se os parâmetros que forneceram os melhores resultados no treinamento, mostrados na Tabela 5.3. A Tabela 5.6 mostra os resultados numéricos para os testes, em que os detectores foram combinados da seguinte forma:

- Detectores 1, 2 e 3 (D1D2D3);

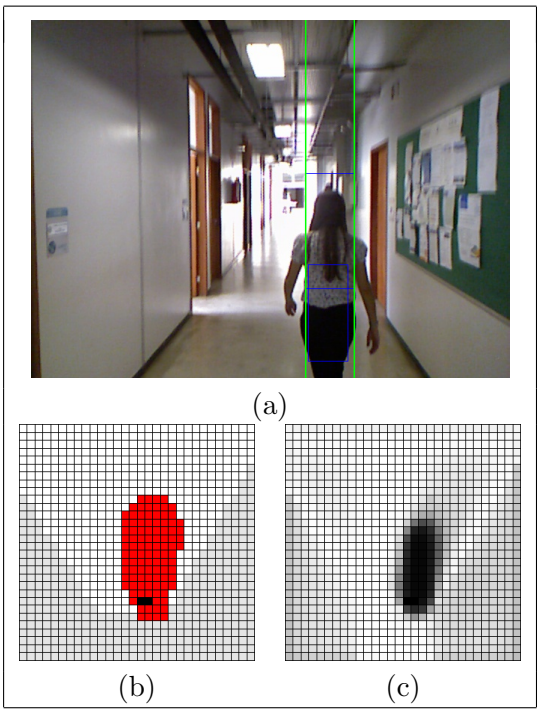


Figura 5.15: Detecção de pessoa em imagem mapeada para a grade com grande erro de localização. (a) Imagem da Base 2 contendo uma pessoa com os pés oclusos. (b) Detecção da pessoa na imagem (a) pelo Detector 3, marcada pelos retângulos azuis. Para o mapeamento da localização da pessoa para a grade, são consideradas as linhas verticais em verde, gerando as células vermelhas da grade. A localização real da pessoa é aproximada pelas células pretas, de acordo com a análise manual da leitura fornecida pelo laser. (c) Na etapa de atualização, a presença da pessoa causa um aumento da probabilidade de um conjunto maior de células.

Tabela 5.6: Resultados da detecção de pessoas obtidos com os dados de teste da Base 2. Os detectores comparados são: Detector 1 [Bellotto e Hu, 2009], Detector 2 [Spinello e Siegwart, 2008], Detector 3 [Dollar et al., 2014] e metodologia proposta seguida da classificação com as seguintes combinações dos detectores: Detectores 1, 2 e 3 (D1D2D3), Detectores 1 e 2 (D1D2), Detectores 1 e 3 (D1D3) e Detectores 2 e 3 (D2D3).

Detector	Revocação	Precisão	FPCPI	Medida-$F1$
Detector 1	$0,43$	$0,79$	$0,13$	$0,56$
Detector 2	$0,36$	$0,44$	$0,67$	$0,40$
Detector 3	$0,56$	$0,70$	$6,72$	$0,62$
D1D2D3	$0,71$	$0,70$	$4,15$	$0,71$
D1D2	$0,58$	$0,69$	$0,69$	$0,63$
D1D3	$0,70$	$0,72$	$3,56$	$0,71$
D2D3	$0,49$	$0,82$	$4,22$	$0,62$

- Detectores 1 e 2 (D1D2);
- Detectores 1 e 3 (D1D3); e
- Detectores 2 e 3 (D2D3).

A metodologia proposta combinada com a classificação, em comparação com todos os detectores, mostrou um aumento no número de detecções corretas (revocação) e na medida-$F1$, o que significa que o número de falsos negativos foi reduzido e um maior número de pessoas que estavam na cena foram detectadas. No geral, a precisão manteve-se entre os dois maiores valores dos detectores individuais. A única combinação de detectores que não resultou em aumento na revocação e na medida-$F1$ foi D2D3, mas em compensação a precisão ficou muito acima da precisão dos detectores individuais e a medida-$F1$ foi mantida igual à maior dos detectores. O maior valor de FPCPI apresentado foi do Detector 3, indicando que esse detector apresenta erros no posicionamento das pessoas, como esperado. Estes resultados estão de acordo com a ideia de que a fusão torna a detecção de pessoas mais confiável que a detecção usando apenas detectores individuais. Os resulta-

dos com as combinações de detectores mostraram que é possível realizar a fusão de detectores distintos mesmo que alguns desses detectores apresente resultados ruins, pois com a metodologia proposta as detecções corretas são aproveitadas adequadamente.

Na maior parte das situações a metodologia proposta melhora o resultados dos detectores individuais, mas por outro lado em outras situações isso pode não ocorrer. Por exemplo, em situações em que alguns dos detectores usados deixam de encontrar muitas pessoas ou que detectam muitos falsos positivos. Nestes casos, a metodologia proposta apresentará resultados insatisfatórios pois a probabilidade de existir pessoas é calculada de acordo com a confiança atribuída aos detectores e, portanto, se os detectores falharem em várias iterações seguidas as informações incorretas obtidas irão refletir nos resultados da metodologia.

Como desvantagem da implementação da metodologia proposta, pode-se citar o tempo de processamento. Como a implementação da metodologia foi realizada no Matlab e os experimentos executados em um computador Intel Core2 Duo $1,8$ GHz, o tempo de processamento não é adequado para execução em tempo real, sendo aproximadamente igual a $1,78$ segundos por iteração para a grade com 30×30 células. Esse tempo não inclui o processamento dos detectores utilizados (no mesmo computador, a detecção nos dados do laser usando o algoritmo de Spinello e Siegwart [2008] leva $0,001$ segundos para cada leitura do laser em uma implementação em C++ e a detecção nas imagens usando o algoritmo de Dollar et al. [2014] leva em média $0,28$ segundos cada imagem no Matlab e $0,03$ segundos na CPU dos autores do detector). Para uma implementação em tempo real, a codificação em linguagem C e a paralelização de algumas etapas da metodologia devem ser consideradas. Em relação à complexidade computacional de tempo, o

custo é $\mathcal{O}(K)$, onde K é o número de células da grade, considerando para este cálculo todas as operações realizadas a partir do momento em que os dados dos detectores já foram carregados (ou seja, não considera o custo dos detectores).

Outra desvantagem é que, em aplicações como navegação de robôs, a implementação utilizando apenas uma grade no mesmo plano que o laser não é adequada para ambientes que possuem aclives ou declives, pois não acompanha o relevo do chão pelo qual o robô vai passar e também não fornece informações sobre pessoas em outros planos. Além disso, a utilização de câmera monocular não permitirá uma aproximação adequada da distância das pessoas por meio do método utilizado na implementação (Seção 4.4), que se restringe a ambientes planos. Essas informações podem ser adquiridas, por exemplo, utilizando sensores a laser com múltiplas camadas ou câmeras de visão estéreo. A implementação, neste caso, deverá contar com múltiplas grades abrangendo outros planos, porém utilizando a mesma metodologia proposta para calcular as probabilidades das células.

Os resultados experimentais desta subseção indicam que, mesmo com o movimento dos sensores, a metodologia proposta provê resultados confiáveis que podem ser usados em diversas aplicações, como por exemplo o rastreamento de pessoas. Para ilustrar um exemplo de uma aplicação da metodologia, a grade final de probabilidades *a posteriori* seguida da classificação baseada em *blobs* foi utilizada para rastrear uma pessoa movendo-se dentro do campo de visão dos sensores do robô, como mostrado a seguir.

5.2.5 Estudo de caso: rastreamento

O rastreamento de pessoas consiste em estimar a trajetória das pessoas ao longo do tempo. Neste estudo de caso, o rastreamento de uma pessoa foi

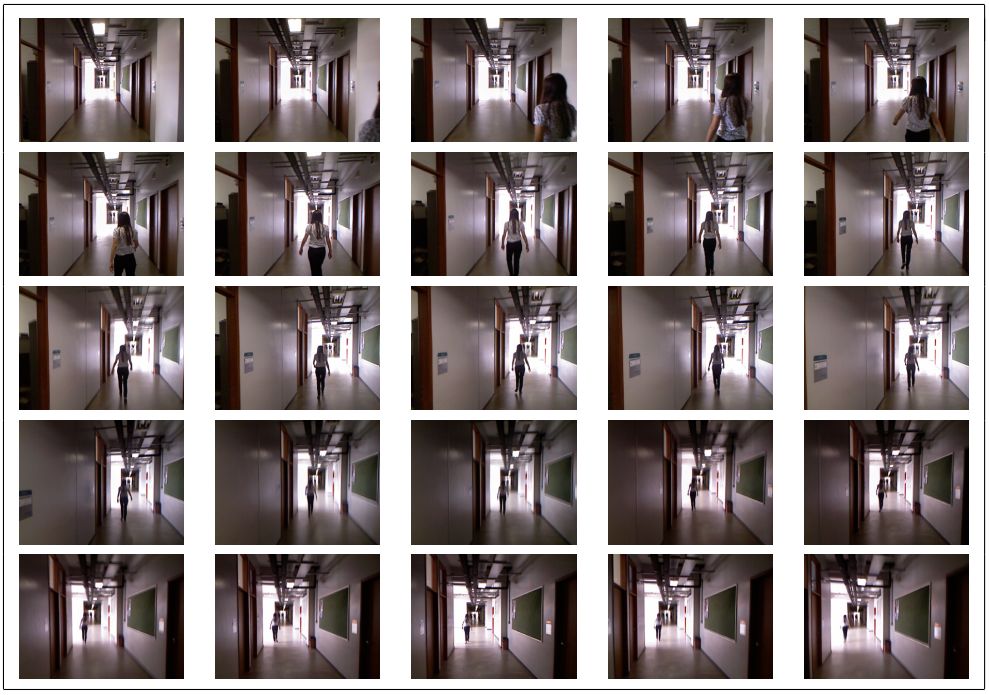


Figura 5.16: Sequência de imagens obtidas durante o rastreamento da Figura 5.17(a) (é mostrada uma imagem a cada cinco iterações). Da esquerda para a direita e de cima para baixo, a pessoa rastreada atravessa a grade iniciando em uma posição que está fora do alcance da câmera. Nas primeiras imagens, a pessoa não pode ser vista pela câmera mas ela está dentro do campo de visão do laser.

realizado no referencial do robô utilizando uma sequência de informações dos Detectores 1, 2 e 3, em um ambiente retratado pelas imagens da Figura 5.16. A pessoa foi rastreada durante vários instantes de tempo, mesmo quando ela não foi detectada por nenhum dos sensores.

A pessoa rastreada inicia seu movimento e passa a ser representada na grade assim que ela entra no campo de visão do laser. Ela continua seu movimento seguindo em frente no corredor e em determinado instante de tempo entra no campo de visão da câmera. O rastreamento iniciou com a seleção manual do *blob* que representa a pessoa escolhida e em seguida os próximos *blobs* foram selecionados automaticamente na grade da classificação

de *blobs*. O critério de seleção do *blob* seguinte é baseado em sua posição em relação ao centroide do *blob* selecionado no instante anterior. Caso existam mais de um *blob* próximo ao centroide, o *blob* com a maior probabilidade *a posteriori* é escolhido. O centroide neste caso foi selecionado como sendo a célula cujas coordenadas são a média das coordenadas das células com a maior probabilidade do *blob*. Se não houver nenhum *blob* próximo ao centroide do *blob* anterior, é escolhido o *blob* de maior probabilidade que se sobrepõe a qualquer célula do *blob* selecionado no instante anterior ou então, como última opção, é escolhido um *blob* na vizinhança-8 de alguma célula do *blob* anterior.

A Figura 5.17(a) mostra o resultado do rastreamento durante 111 instantes de tempo consecutivos. Nessa figura são mostrados os centroides dos *blobs* selecionados (em preto). Como comparação, a Figura 5.17(b) mostra as células (em vermelho) por onde a pessoa passou, de acordo com o *ground-truth* obtido manualmente a partir dos dados do laser. Nota-se que, em relação ao *ground-truth*, o rastreamento foi bem sucedido. Durante o rastreamento, os Detectores 1, 2 e 3 falharam simultaneamente em detectar a pessoa em 6 instantes de tempo. O Detector 1 encontrou a pessoa em 77 % das iterações, enquanto que os Detectores 2 e 3 detectaram a pessoa 30 % e 79 % do tempo, respectivamente, o que mostrou a robustez da metodologia proposta nesta situação, que manteve a detecção da pessoa durante sua trajetória mesmo com falhas nos detectores.

5.3 Discussão

A fusão dos detectores testados usando a metodologia proposta mostrou as vantagens da abordagem em relação a detectores individuais. Nos exem-

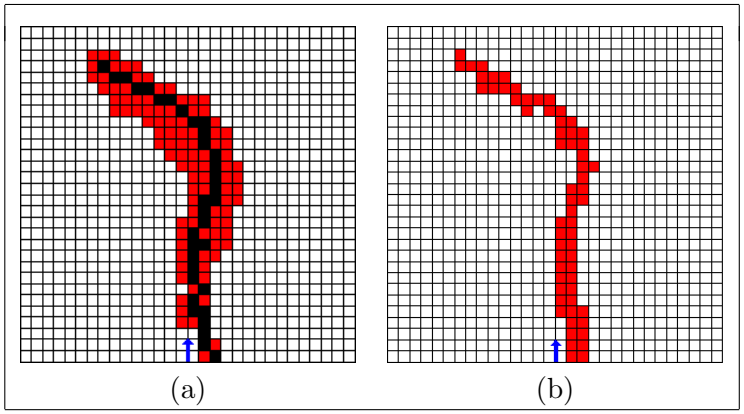


Figura 5.17: Resultados do rastreamento de uma pessoa na grade. A seta indica o início do rastreamento. (a) As células pretas representam os centroides dos *blobs* selecionados (células vermelhas). (b) *Ground-truth* obtido manualmente dos dados do laser.

plos mostrados na Figura 5.11, é observado que o resultado final apresentado na Figura 5.11(r) é (i) mais completo que a informação obtida de cada um dos detectores, que falharam simultaneamente na detecção de uma das pessoas da cena, (ii) mais preciso em relação à detecção, uma vez que a detecção incorreta fornecida por um dos detectores não foi incorporada ao resultado final (Figure 5.11(o)), e (iii) mais confiável em relação à localização das pessoas, visto que o resultado obtido pela câmera foi refinado pela informação do laser.

A metodologia proposta possibilita o uso de diversas combinações de detectores de pessoas e é flexível em relação aos sensores utilizados, permitindo sua utilização em vários contextos. Outra flexibilidade da metodologia é em relação ao modelo de movimento das pessoas, que mostrou-se importante para os resultados finais. No modelo de movimento utilizado, considera-se que o movimento das pessoas têm uma probabilidade uniforme em todas as direções. Se a direção do movimento é controlada pela presença de um corredor, por exemplo, esta informação pode ser usada para determinar um

modelo mais preciso.

Em geral, os resultados experimentais usando a grade de ocupação semântica para representar localmente o ambiente mostraram que a metodologia proposta apresenta vantagens em relação aos detectores de pessoas individuais, demonstrando que é possível obter um maior número de detecções de pessoas na cena mantendo baixo o número de falsos alarmes. Os trabalhos encontrados na literatura que utilizam laser e câmera para detecção de pessoas possuem características distintas que não permitem uma comparação direta dos resultados, por exemplo devido a utilização de bases de dados diferentes.

Capítulo 6

Considerações finais

A detecção de pessoas em ambientes não controlados de forma eficiente e robusta é uma tarefa desafiadora almejada em várias aplicações de robótica, tais como navegação, sistemas de apoio ao motorista, resgate e busca de pessoas, interação humano-robô, etc. Neste contexto, este trabalho de doutorado propôs uma metodologia bayesiana para a fusão de detectores de pessoas, com o objetivo de explorar as vantagens de diversos detectores e produzir um resultado mais confiável e robusto. A metodologia proposta leva em consideração o movimento das pessoas e dos sensores em duas etapas distintas de predição e combina as informações de múltiplos detectores de pessoas, que além de diversas podem ser conflitantes, de forma adequada na etapa de atualização. Estas informações de alto-nível são combinadas de modo que, quando mais de um detector indica a presença de uma pessoa em uma dada posição, a confiança da detecção é maior que a confiança no caso de apenas um detector detectar uma determinada pessoa.

Experimentos utilizando uma grade de ocupação semântica e detectores distintos permitiram a construção de uma grade local com informações sobre a presença de pessoas em um ambiente dinâmico. A metodologia proposta foi testada com quatro sensores (câmera, Kinect e dois tipos de laser planares) e com cinco detectores distintos, mostrando a escalabilidade em relação

ao número de detectores e a versatilidade em relação aos tipos de sensores utilizados. Os resultados mostraram que a metodologia proposta apresenta vantagens em relação ao estado da arte de detecção de pessoas, demonstrando que é possível obter um maior número de detecções de pessoas na cena mantendo baixo o número de falsos alarmes levando-se em consideração a confiança de cada um dos detectores utilizados e informação de instantes passados. Além disso, os resultados demonstraram que os objetivos da tese foram alcançados, ou seja, dados os resultados de N detectores, a metodologia proposta para a fusão obteve informações mais completas sobre as pessoas presentes no ambiente e resultados mais confiáveis em relação à sua localização que os resultados dos detectores individuais, além de ser robusta a falhas dos detectores, maximizando o número de detecções e mantendo o número de falsos alarmes baixo.

Os detectores de pessoas do estado da arte mostram que ainda existem limitações para a utilização em aplicações que possam envolver riscos às pessoas presentes no ambiente, pois a precisão e número de falsos alarmes da detecção de pessoas indicam que existe a possibilidade de pessoas não serem detectadas pelo robô e também do robô detectar a presença de pessoas que não existem. A metodologia proposta nesta tese formalizou uma maneira de combinar detectores que trabalham com dados de naturezas distintas, possibilitando um ganho nos resultados dos detectores individuais com informações mais precisas. A utilização de uma diversidade de detectores de pessoas mais confiáveis consequentemente resultará em melhores resultados.

6.1 Trabalhos Futuros

Além da utilização de detectores mais confiáveis, outras questões devem ser consideradas para a construção de sistemas de detecção de pessoas mais

eficientes:

- Modelo de movimento das pessoas: o modelo de movimento contribui para a predição da posição das pessoas no próximo instante de tempo e deve incorporar ao máximo a diversidade do comportamento das pessoas bem como a tendência de comportamentos típicos. Por exemplo, em corredores a tendência das pessoas é caminhar na direção paralela ao corredor. Um modelo de pessoas adaptativo que leva em consideração o ambiente em que o robô se encontra pode ser proposto futuramente, preferencialmente não necessitando de informações de mapas globais, que podem não estar disponíveis, mas utilizando informações dos próprios sensores do robô para realizar uma aproximação do mapa local.

- Detectores com informações mais ricas sobre o ambiente: a utilização de detectores baseados em sensores que fornecem informações distintas e complementares possibilita a percepção de características diferentes do ambiente, mesmo dentro de um mesmo campo de visão. Por exemplo, os detectores baseados em laser e os detectores baseados em imagens utilizados nos experimentos. Futuramente podem ser usados outros detectores que trabalhem com outros tipos de características, sendo baseados em outros tipos de sensores. Além disso, informações sobre outros planos do ambiente devem ser consideradas por meio de detectores baseados em sensores que forneçam informações 3D, por exemplo câmeras de visão estéreo e sensores a laser 3D, como o Velodyne.

- Custo e tecnologia dos sensores: o custo dos sensores pode ser um fator limitante para que sistemas de detecção de pessoas possam ser utilizados em aplicações comerciais. Futuramente, avanços na tecnologia podem permitir novos tipos de sensores e redução dos custos. Com os

sensores da atualidade, há questões limitantes além do custo, como a interferência no uso em massa nos sensores ativos, grande quantidade de dados produzidos, como nos sensores 3D, e restrições da tecnologia, por exemplo dependência de condições de iluminação, efeito de reflexões ou de baixa refletividade de alguns materiais, incidência direta da luz do sol, alcance, frequência de amostragem, etc. Para ilustrar uma destas situações, a influência da distância das pessoas na percepção de um laser 2D de uma camada a uma altura equivalente aos joelhos de uma pessoa é mostrada no gráfico da Figura 6.1. O gráfico mostra no eixo x a distância da pessoa em relação ao laser e no eixo y a distância entre os raios do laser, para uma resolução angular de $0,125$ graus. O alcance e a resolução angular do laser afetam a qualidade das detecções, pois à medida que a pessoa se afasta, a distância entre os raios passa a ser maior que a largura de suas pernas (por exemplo acima de 40 metros a distância entre os raios fica maior que $0,2$ metros) e desta forma os raios do laser podem não atingir a pessoa e consequentemente não haverá detecção. Todas essas questões devem ser levadas em consideração para determinar as limitações dos detectores de pessoas nas aplicações.

- Custo de processamento: para implementação da metodologia proposta em aplicações de tempo real devem ser consideradas linguagens mais eficientes e a paralelização do processamento. É importante notar que, quanto maior for a velocidade do robô, maior deve ser a frequência com que os detectores fornecem informações para a etapa de atualização, pois nesta situação a incerteza na localização das pessoas aumenta consideravelmente pelo fato de que o robô percorreu uma distância maior. O processamento a taxas mais altas será possível com proces-

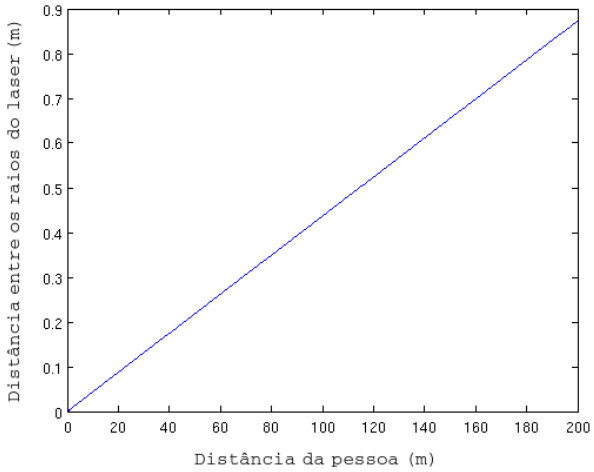


Figura 6.1: Distância da pessoa em relação ao laser × Distância entre os raios do laser, para uma resolução angular de 0, 125 graus.

sadores mais rápidos e também com o uso do processamento paralelo de GPUs (*Graphics Processing Unit*), considerando que algumas etapas da abordagem podem ser paralelizáveis devido à independência do cálculo das probabilidades das células da grade. Trabalhos anteriores que usaram GPUs conseguiram diminuir consideravelmente o tempo de processamento das abordagens em relação ao uso de computadores sem GPU. Por exemplo, no trabalho de Baig et al. [2014] um dos módulos da abordagem, quando implementado em GPU teve seu tempo de processamento reduzido em aproximadamente 32%. Já Yoder et al. [2014] conseguiu uma redução próxima de 95% ao tirar proveito da estrutura paralela do BOF (*Bayesian Occupancy Filter*), derivado de grades de ocupação. O trabalho de Yguel et al. [2006] também usa uma GPU para realizar transformações de sistemas de coordenadas e a fusão entre as grades de ocupação, que são construídas utilizando vários sensores laser e relata que o uso de GPU permite a fusão de grade de ocupação para 50 sensores simultaneamente a uma taxa equivalente

à de leitura do sensor. Além do processamento paralelo em GPU, os códigos implementados no Matlab podem ser convertidos futuramente em versões mais eficientes em linguagem C [Zaranek et al., 2013].

Referências Bibliográficas

Adarve, J., Perrollaz, M., Makris, A., e Laugier, C. (2012). Computing occupancy grids from multiple sensors using linear opinion pools. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 4074–4079.

Aguirre, L. (2007). *Introdução à Identificação de Sistemas – Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Editora UFMG.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd ed.

Antonini, G., Bierlaire, M., e Weber, M. (2006). Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687.

Antunes, M., Barreto, J., Premebida, C., e Nunes, U. (2012). Can stereo vision replace a laser rangefinder? In *Proceeings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5183–5190.

Araujo, A. R., Caminhas, D. D., e Pereira, G. A. S. (2015). An architecture for navigation of service robots in human-populated office-like environments. In *Proceedings of the IFAC Symposium on Robot Control (submitted)*.

Araújo, R. L., Lacerda, V., Hernandes, A., Mendonca, A., e Becker, M. (2011). Classificação de pedestres usando câmera e sensor lidar. In *Anais do Simpósio Brasileiro de Automação Inteligente*, pp. 416–420.

Baig, Q., Perrollaz, M., e Laugier, C. (2014). A robust motion detection technique for dynamic environment monitoring: A framework for grid-based monitoring of the dynamic environment. *IEEE Robotics Automation Magazine*, 21(1):40–48.

Batista, N. C. e Pereira, G. A. S. (2013). Avaliação de técnicas para detecção de pedestres usando laser e câmera visando a fusão sensorial. In *Anais do Simpósio Brasileiro de Automação Inteligente*.

Bellotto, N. e Hu, H. (2009). Multisensor-based human detection and tracking for mobile service robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(1):167–181.

Benenson, R., Omran, M., Hosang, J., e Schiele, B. (2015). Ten years of pedestrian detection, what have we learned? In Agapito, L., Bronstein, M. M., e Rother, C., editors, *Computer Vision - ECCV 2014 Workshops*, *Lecture Notes in Computer Science*, volume 8926, pp. 613–627. Springer International Publishing.

Bengler, K., Dietmayer, K., Farber, B., Maurer, M., Stiller, C., e Winner, H. (2014). Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22.

Bertozzi, M., Broggi, A., Cellario, M., Fascioli, A., Lombardi, P., e Porta, M. (2002). Artificial vision in road vehicles. In *Proceedings of the IEEE*, volume 90, pp. 1258–1271.

Bertozzi, M., Broggi, A., Fascioli, A., Tibaldi, A., Chapuis, R., e Chausse, F. (2004). Pedestrian localization and tracking system with kalman filtering. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 584–589. IEEE.

Biresev, N. (2012). Semantic mapping using object-class segmentation of RGB-D images. Master of science thesis, University of Bonn.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Bota, S. e Nedesvchi, S. (2008). Multi-feature walking pedestrians detection for driving assistance systems. *IET Transactions on Intelligent Transport Systems*, 2(2):92–104.

Bouguet, J.-Y. (2010). Camera calibration toolbox for matlab. Disponível em: `http://www.vision.caltech.edu/bouguetj/calib_doc/`. Acesso em: 03 de junho de 2013.

Bouzouraa, M. e Hofmann, U. (2010). Fusion of occupancy grid mapping and model based object tracking for driver assistance systems using laser and radar sensors. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 294–300.

Braga, A. d. P., Carvalho, A. C. P. d. L. F. d., e Ludermir, T. B. (2007). *Redes neurais artificiais: teoria e aplicações*. LTC, Rio de Janeiro, 2 ed.

Broggi, A., Cerri, P., Ghidoni, S., Grisleri, P., e Jung, H. G. (2009). A new approach to urban pedestrian detection for automatic braking. *IEEE Transactions on Intelligent Transportation Systems*, 10(4):594–605.

Bu, F. e Chan, C.-Y. (2005). Pedestrian detection in transit bus application: sensing technologies and safety solutions. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 100–105.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Ceccarelli, M. (2011). Problems and issues for service robots in new applications. *International Journal of Social Robotics*, 3(3):299–312.

Cho, H., Seo, Y.-W., Vijaya Kumar, B., e Rajkumar, R. (2014). A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1836–1843.

Choset, H., Lynch, K. M., Hutchinson, S., Kantor, G. A., Burgard, W., Kavraki, L. E., e Thrun, S. (2005). *Principles of Robot Motion: Theory, Algorithms, and Implementations*. MIT Press, Cambridge, MA.

Cui, J., Zha, H., Zhao, H., e Shibasaki, R. (2005). Tracking multiple people using laser and vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2116–2121.

Daamen, W. e Hoogendoorn, S. (2007). Free speed distributions - based on empirical data in different traffic conditions. In *Pedestrian and Evacuation Dynamics 2005*, pp. 13–25. Springer Berlin Heidelberg.

Dalal, N. e Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893.

Devroye, L., Györfi, L., e Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer New York.

Discant, A., Rogozan, A., Rusu, C., e Bensrhair, A. (2007). Sensors for obstacle detection - a survey. In *30th International Spring Seminar on Electronics Technology*, pp. 100–105.

Dollar, P., Appel, R., Belongie, S., e Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545.

Dollar, P., Wojek, C., Schiele, B., e Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

Elfes, A. (1990). Occupancy grids: A stochastic spatial representation for active robot perception. In *Proceedings of the Conference of Uncertainty in Artificial Intelligence*, pp. 136–146. AUAI Press.

Elguebaly, T. e Bouguila, N. (2011). A nonparametric bayesian approach for enhanced pedestrian detection and foreground segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 21–26.

Filliat, D., Battesti, E., Bazeille, S., Duceux, G., Gepperth, A., Harrath, L., Jebari, I., Pereira, R., Tapus, A., Meyer, C., Ieng, S.-H., Benosman, R., Cizeron, E., Mamanna, J.-C., e Pothier, B. (2012). Rgbd object recognition and visual texture classification for indoor semantic mapping. In *Proceedings of the IEEE International Conference on Technologies for Practical Robot Applications*, pp. 127–132.

Ge, J., Luo, Y., e Tei, G. (2009). Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems*, 10(2):283–298.

Gepperth, A., Sattarov, E., Heisele, B., e Rodriguez Flores, S. (2014). Robust visual pedestrian detection by tight coupling to tracking. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 1935–1940.

Geronimo, D., Lopez, A., Sappa, A., e Graf, T. (2010). Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258.

Geronimo, D. e Lopez, A. M. (2014). *Vision-based Pedestrian Protection Systems for Intelligent Vehicles*. Springer-Verlag New York.

Gidel, S., Blanc, C., Chateau, T., Checchin, P., e Trassoudaine, L. (2009). Non-parametric laser and video data fusion: Application to pedestrian detection in urban environment. In *Proceedings of the International Conference on Information Fusion*, pp. 626–632.

Gonzalez, R. C. e Woods, R. E. (2003). *Processamento de Imagens Digitais*. Editora Edgard Blucher, 1st ed.

Haykin, S. (2001). *Redes Neurais*. Bookman Companhia Ed.

Hofmann, M., Kaiser, M., Aliakbarpour, H., e Rigoll, G. (2011). Fusion of multi-modal sensors in a voxel occupancy grid for tracking and behaviour analysis. In *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*.

Hogenboom, M. (2013). Secret of Usain Bolt's speed unveiled. `http://www.bbc.co.uk/news/science-environment-23462815`. Acesso em: 06 de novembro de 2014.

Huerta, I., Ferrer, G., Herrero, F., Prati, A., e Sanfeliu, A. (2014). Multimodal feedback fusion of laser, image and temporal information. In *Proceedings of the International Conference on Distributed Smart Cameras*, pp. 25:1–25:6, New York, USA. ACM.

Jebari, I., Bazeille, S., Battesti, E., Tekaya, H., Klein, M., Tapus, A., Filliat, D., Meyer, C., Ieng, S.-H., Benosman, R., Cizeron, E., Mamanna, J.-C., e Pothier, B. (2011). Multi-sensor semantic mapping and exploration of indoor environments. In *Proceedings of the IEEE Conference on Technologies for Practical Robot Applications*, pp. 151–156.

Kassir, A. e Peynot, T. (2010). Reliable automatic camera-laser calibration. In *Proceedings of the Australasian Conference on Robotics & Automation*, p. 10.

Keller, C. e Gavrila, D. (2014). Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506.

Kim, S., Guy, S. J., Liu, W., Wilkie, D., Lau, R. W., Lin, M. C., e Manocha, D. (2015). Brvo: Predicting pedestrian trajectories using velocity-space reasoning. *International Journal of Robotics Research*, 34(2):201–217.

Kooij, J., Schneider, N., e Gavrila, D. (2014). Analysis of pedestrian dynamics from a vehicle perspective. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 1445–1450.

Lima, D. A. e Pereira, G. A. S. (2010). Um sistema de visão estéreo para navegação de um carro autônomo em ambientes com obstáculos. In *Anais do XVIII Congresso Brasileiro de Automática*, pp. 224–231.

Linzmeier, D. (2004). Pedestrian detection with thermopiles using an occupancy grid. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 1063–1068.

Linzmeier, D., Mekhaiel, M., Dickmann, J., e Dietmayer, K. C. J. (2004). Pedestrian detection with thermopiles using an occupancy grid. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pp. 1063–1068.

Liu, Z., Chen, D., e von Wichert, G. (2012). 2D semantic mapping on occupancy grids. In *Proceedings of the German Conference on Robotics*, pp. 1–6.

Liu, Z. e von Wichert, G. (2014). Extracting semantic indoor maps from occupancy grids. *Robotics and Autonomous Systems*, 62(5):663–674.

Lorena, A. C. e de Carvalho, A. C. P. L. F. (2007). Uma introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Lu, Z., Hu, Z., Uchimura, K., Kubota, H., e Ono, M. (2008). Sensor fusion with occupancy fusion map for pedestrian detection in outdoor environment. In *Proceedings of the IEEE International Conference on Vehicular Electronics and Safety*, pp. 199–204.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., e Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84(8):2200–2207.

Messeder, D. (2014). Garagem CARPLACE \# 6: Ka+ encara Grand Siena, Prisma e Logan. `http://carplace.uol.com.br/garagem-carplace-6-ka-encara-grand-siena-prisma-e-logan/`. Acesso em: 18 de maio de 2015.

Monteiro, G., Premebida, C., Peixoto, P., e Nunes, U. (2006). Tracking and classification of dynamic obstacles using laser range finder and vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Moras, J., Cherfaoui, V., e Bonnifait, P. (2011). Credibilist occupancy grids for vehicle perception in dynamic environments. In *Proceedings IEEE International Conference on Robotics and Automation*, pp. 84–89.

Mozos, O. M., Kurazume, R., e Hasegawa, T. (2010). Multi-part people detection using 2D range data. *International Journal of Social Robotics*, 2(1):31–40.

Ngako Pangop, L., Chausse, F., Chapuis, R., e Cornou, S. (2008). Asynchronous bayesian algorithm for object classification: Application to pedestrian detection in urban areas. In *Proceedings of the International Conference on Information Fusion*, pp. 1–7.

Ngako Pangop, L., Chausse, F., Cornou, S., e Chapuis, R. (2007). Feature-based multisensor fusion using bayes formula for pedestrian classification in outdoor environments. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 62–67.

Nilsson, N. J. (1998). *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, Oxford.

Nüchter, A. e Hertzberg, J. (2008). Towards semantic maps for mobile robots. *Journal of Robotics and Autonomous Systems*, 56(11):915–926.

Oliveira, L., Nunes, U., e Peixoto, P. (2010a). On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 11(1):16–27.

Oliveira, L., Nunes, U., Peixoto, P., Silva, M., e Moita, F. (2010b). Semantic fusion of laser and vision in pedestrian detection. *Pattern Recognition*, 43:3648–3659.

Papoulis, A. e Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*. Mc-Graw Hill, 4th ed.

Pedrini, H. e Schwartz, W. R. (2008). *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações*. Thomson Learning.

Pereira, F. G., Vassallo, R. F., e Salles, E. O. T. (2013). Human-robot interaction and cooperation through people detection and gesture recognition. *Journal of Control, Automation and Electrical Systems*, 24(3):187–198.

Premebida, C., Ludwig, O., e Nunes, U. (2009). Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, 26(9):696–711.

Premebida, C. e Nunes, U. J. C. (2013). Fusing lidar, camera and semantic information: A context-based approach for pedestrian detection. *The International Journal of Robotics Research*, 32(3):371–384.

Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., Wheeler, R., e Ng, A. Y. (2009). Ros: an open-source robot operating system. In *Proceedings of the ICRA Workshop on Open Source Software*, volume 3.

Ribo, M. e Pinz, A. (2001). A comparison of three uncertainty calculi for building sonar-based occupancy grids. *Robotics and Autonomous Systems*, 35(3-4):201–209.

Ros, J. e Mekhnacha, K. (2009). Multi-sensor human tracking with the bayesian occupancy filter. In *Proceedings of the International Conference on Digital Signal Processing*, pp. 1–8.

Sabbagh, V., Freitas, E., Castro, G., Santos, M., Baleeiro, M., Silva, T., Iscold, P., Torres, L., e Pereira, G. (2010). Desenvolvimento de um sistema de controle para um carro de passeio autônomo. In *Anais do XVIII Congresso Brasileiro de Automática*, pp. 928–933.

Saxena, A., Schulte, J., e Ng, A. Y. (2007). Depth estimation using monocular and stereo cues. In *Proceedings of the International Joint Conference on Artifical Intelligence*, pp. 2197–2203. Morgan Kaufmann Publishers Inc.

Schapire, R. E. (2013). Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*.

Schneider, N. e Gavrila, D. (2013). Pedestrian path prediction with recursive bayesian filters: A comparative study. In Weickert, J., Hein, M., e Schiele, B., editors, *Pattern Recognition*, *Lecture Notes in Computer Science*, volume 8142, pp. 174–183. Springer Berlin Heidelberg.

Shi, L., Kodagoda, S., e Dissanayake, G. (2010). Environment classification and semantic grid map building based on laser range finder data. In *Proceedings of the IROS Workshop on Semantic Mapping and Autonomous Knowledge Acquisition*, pp. 1–6.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., e Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 1297–1304. IEEE Computer Society.

SICK (2006). Technical description for the lms200/211/221/291 laser measurement systems. Technical report, SICK AG Waldkirch, Germany.

Spinello, L. e Siegwart, R. (2008). Human detection using multimodal and multidimensional features. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3264–3269.

Stein, G., Mano, O., e Shashua, A. (2003). Vision-based acc with a single camera: bounds on range and range rate accuracy. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 120–125.

Tay, M., Mekhnacha, K., Yguel, M., Coué, C., Pradalier, C., Laugier, C., Fraichard, T., e Bessière, P. (2008). The bayesian occupation filter. In *Probabilistic Reasoning and Decision Making in Sensory-Motor Systems*, *Springer Tracts in Advanced Robotics*, volume 46, pp. 77–98. Springer Berlin Heidelberg.

Teixeira, T., Dublon, G., e Savvides, A. (2010). A survey of human-sensing: Methods for detecting presence, count, location, track, and identity. Technical Report ENALAB Technical Report 09-2010, Yale University, Pittsburgh, PA.

Theodoridis, S. e Koutroumbas, K. (2008). *Pattern Recognition*. Elsevier Science.

Thrun, S. (2003). Learning occupancy grid maps with forward sensor models. *Autonomous Robots*, 15(2):111–127.

Thrun, S., Burgard, W., e Fox, D. (2005). *Probabilistic Robotics*. The MIT Press, Cambridge, MA.

Trucco, E. e Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River, NJ, USA.

Unnikrishnan, R. e Hebert, M. (2005). Fast extrinsic calibration of a laser rangefinder to a camera. Technical Report CMU-RI-TR-05-09, Robotics Institute, Pittsburgh, PA.

Utasi, A. e Benedek, C. (2013). A bayesian approach on people localization in multicamera systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1):105–115.

Varga, R., Vesa, A., Jeong, P., e Nedevschi, S. (2014). Real-time pedestrian detection in urban scenarios. In *Proceedings of the IEEE International*

Conference on Intelligent Computer Communication and Processing, pp. 113–118.

Varvadoukas, T., Giotis, I., e Konstantopoulos, S. (2012). Detecting human patterns in laser range data. In *Proceedings of the European Conference on AI*, volume 242, pp. 804–809.

Viola, P. e Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 511–518.

Vu, T.-D., Aycard, O., e Tango, F. (2014). Object perception for intelligent vehicle applications: A multi-sensor fusion approach. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 774–780.

Wang, T. e Chen, Q. (2011). Object semantic map representation for indoor mobile robots. In *Proceedings of the International Conference on System Science and Engineering*, pp. 309–313.

Weinrich, C., Wengefeld, T., Schroeter, C., e Gross, H.-M. (2014). People detection and distinction of their walking aids in 2D laser range data based on generic distance-invariant features. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pp. 767–773.

Wolf, D. e Sukhatme, G. (2008). Semantic mapping using mobile robots. *IEEE Transactions on Robotics,*, 24(2):245–258.

Wu, B., Liang, J., Ye, Q., Han, Z., e Jiao, J. (2011). Fast pedestrian detection with laser and image data fusion. In *Proceedings of the International Conference on Image and Graphics*, pp. 605 –608.

Yguel, M., Aycard, O., e Laugier, C. (2006). Efficient gpu-based construction of occupancy girds using several laser range-finders. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 105–110.

Yoder, J.-D., Perrollaz, M., Paromtchik, I., Mao, Y., e Laugier, C. (2014). Experiments in Vision-Laser Fusion using the Bayesian Occupancy Filter. In *Experimental Robotics*, 79(1):899-907.

Zaranek, S. W., Chou, B., Sharma, G., e Zarrinkoub, H. (2013). Accelerating MATLAB Algorithms and Applications. `http://www.mathworks.com/company/newsletters/articles/`

`accelerating-matlab-algorithms-and-applications.html`. Acesso em: 02 de junho de 2015.

Zhang, Q. e Pless, R. (2004). Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *Proceedings of the International Conference on Intelligent Robots and Systems*, volume 3, pp. 2301–2306.

Zhang, S., Benenson, R., e Schiele, B. (2015). Filtered channel features for pedestrian detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

Zhao, T. e Nevatia, R. (2003). Bayesian human segmentation in crowded situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 459–466.

Zhi-yu, X., Ji-lin, L., Wei-kang, G., e Yuan, T. (2001). Obstacle detection by alv using two 2d laser range finders. *Journal of Zhejiang University Science*, 2(4):388–394.