Douglas Almonfrey

Serviço Flexível de Detecção de Seres Humanos para Espaços Inteligentes Baseados em Redes de Câmeras

Brasil
2018

Douglas Almonfrey

Serviço Flexível de Detecção de Seres Humanos para Espaços Inteligentes Baseados em Redes de Câmeras

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, do Centro Tecnológico da Universidade Federal do Espírito Santo, para a obtenção do Grau de Doutor em Engenharia Elétrica - Robótica e Automação Inteligente (RO).

Universidade Federal do Espírito Santo
Centro Tecnológico
Programa de Pós-graduação em Engenharia Elétrica

Orientadora: Prof[a]. Dr[a]. Raquel Frizera Vassallo
Coorientador: Prof. Dr. Evandro Ottoni Teatini Salles

Brasil
2018

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

A452s

Almonfrey, Douglas, 1985-
Serviço flexível de detecção de seres humanos para espaços inteligentes baseados em redes de câmeras / Douglas Almonfrey. – 2018.
107 f. : il.

Orientador: Raquel Frizera Vassallo.
Coorientador: Evandro Ottoni Teatini Salles.
Tese (Doutorado em Engenharia Elétrica) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Seres humanos – Detecção. 2. Pedestres – Detecção. 3. Ambiente inteligente. 4. Robótica. 5. Arquitetura orientada a serviços (Computador). 6. Rede de câmeras. I. Vassallo, Raquel Frizera. II. Salles, Evandro Ottoni Teatini. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 621.3

Elaborada por Sandra Mara Borges Campos – CRB-6 ES-000593/O

Douglas Almonfrey

Serviço Flexível de Detecção de Seres Humanos para Espaços Inteligentes Baseados em Redes de Câmeras

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, do Centro Tecnológico da Universidade Federal do Espírito Santo, para a obtenção do Grau de Doutor em Engenharia Elétrica - Robótica e Automação Inteligente (RO).

Trabalho aprovado. Brasil, 26 de Julho de 2018:

Prof^a. Dr^a. Raquel Frizera Vassallo
Orientadora

Prof. Dr. Evandro Ottoni Teatini Salles
Coorientador

Prof. Dr. Thomas Walter Rauber
Universidade Federal do Espírito Santo - Brasil

Prof. Dr. Patrick Marques Ciarelli
Universidade Federal do Espírito Santo - Brasil

Prof. Dr. João Marques Salomão
Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo - Brasil

Prof. Dr. José Alberto Santos-Victor
Instituto Superior Técnico - Portugal

Brasil
2018

Aos meus avós Pracidino†, Anna†, Acylino† e Rosa, com amor, carinho e saudade.

† ***in memoriam***

Agradecimentos

Eu agradeço primeiramente a Deus e ao nosso senhor Jesus Cristo por, mesmo diante de nossas imperfeições, nos sustentar quando clamamos por força nos momentos de dificuldade.

Eu agradeço aos meus pais Isidoro e Goreth que, com abdicação, amor, carinho e dedicação, possibilitaram que eu alcançasse os títulos de Engenheiro e Mestre. Meus pais foram fundamentais na construção da base que me permitiu concluir esta tese de doutorado. Gostaria também de agradecer, com carinho especial, a minha esposa Letícia pelo apoio e sobretudo paciência, principalmente na reta final deste trabalho. Foram muitas conversas, encorajamentos e abdicações. Te amo Lê.

Eu agradeço especialmente aos meus orientadores Raquel e Evandro, que me acompanharam na jornada até a conclusão deste trabalho, que é, sem dúvida, nosso. Lá se vão dez anos Raquel, desde que te abordei no corredor do CT II para você me orientar no TCC. Obrigado por tudo. Eu agradeço também ao Felippe, Rodolfo e Alexandre pela ajuda no artigo de revista. Sem vocês, tudo teria sido ainda mais difícil. Não poderia deixar de mencionar nominalmente também o Thales, pelo trabalho no processo de anotação das imagens. Agradeço também a professora Mylène por ter, gentilmente, me acolhido na UnB no período que por lá passei.

Nesse momento, faço menção também aos meus demais familiares: minha irmã, afilhado, padrinhos, avós, cunhados(as), sogro(a), tios(as) e primos(as). Muito obrigado pelas orações e presença em determinados momentos em que, às vezes, nem sabiam que estavam ajudando. Agradeço ainda a todos os meus amigos e aos amigos e colegas dos laboratórios VIROS e CISNE pelas discussões e companhia. Mesmo não mencionando um a um, saibam que são muito importantes. Sem dúvida o trabalho em equipe é o caminho.

Agradeço também ao Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo (Ifes), que me financiou por meio do afastamento remunerado para capacitação e aos meus colegas de trabalho que me cobriram (diretamente ou indiretamente) durante esse afastamento. Por fim, mas não menos importante, agradeço a todos os meus professores e a Universidade Federal do Espírito Santo (UFES), na qual executei toda a minha jornada acadêmica.

Com amor e carinho, Douglas Almonfrey.

Resumo

O desenvolvimento de pesquisas voltadas para espaços inteligentes tem sido recorrente na última década. Como uma instância da computação ubíqua, a ideia geral é extrair informação do ambiente e usá-la para interagir e prover serviços para os atores nele presentes. O emprego de sensores é fundamental nessa área e seres humanos são, geralmente, os atores envolvidos. Nesse sentido, nesta tese de doutorado, propõe-se um detector de seres humanos para ser empregado em um espaço inteligente baseado em uma rede de câmeras. O detector é implementado utilizando conceitos de computação em nuvem e arquitetura orientada a serviços (*service-oriented architecture* - SOA). Como principal contribuição deste trabalho, o detector de seres humanos é desenvolvido como um serviço, que é escalável, confiável e paralelizável. É também uma preocupação que o serviço proposto seja flexível, não dependente de *hardware* específico e o menos estruturado possível, atendendo a diferentes aplicações e serviços do espaço inteligente. Uma rede de câmeras, que no cotidiano se encontra normalmente instalada em diferentes ambientes, é empregada para eliminar problemas apresentados por detectores de seres humanos baseados em uma única câmera. De forma a validar a solução desenvolvida, implementam-se três aplicações provas de conceito (PdC) de diferentes tarefas reais do dia a dia. Duas das tarefas apresentadas envolvem a navegação de um robô e demandam a percepção sobre a localização tridimensional dos seres humanos presentes no ambiente. No que diz respeito aos requisitos de tempo e qualidade de detecção, o serviço proposto mostrou-se adequado para interagir com os outros serviços da arquitetura do espaço inteligente, de maneira a completar, de forma bem sucedida, as tarefas relativas a cada aplicação desenvolvida. Como uma contribuição adicional, um procedimento de extração de características, baseado na teoria da análise de componentes independentes (*independent component analysis* - ICA), é proposto como parte de um detector de seres humanos. Testes são conduzidos, em bases de dados públicas, de forma a avaliar o procedimento desenvolvido. A área da detecção de pedestres é empregada como estudo de caso para o desenvolvimento e análise do detector de seres humanos, devido a maturidade dessa área na comunidade científica. O método de extração de características proposto é também utilizado como parte do fluxograma de operação do serviço de detecção de seres humanos desenvolvido. Dessa forma, esse procedimento também é analisado em aplicações de tempo real, no contexto de espaços inteligentes.

Abstract

The topic of intelligent spaces has experienced increasing attention in the last decade. As an instance of the ubiquitous computing paradigm, the general idea is to extract information from the ambient and use it to interact and provide services to the actors present in the environment. The sensory analysis is mandatory in this area and humans are usually the principal actors involved. In this sense, we propose a human detector to be used in an intelligent space based on a multi-camera network. Our human detector is implemented using concepts of cloud computing and service-oriented architecture (SOA). As the main contribution of the present work, the human detector is designed to be a service that is scalable, reliable and parallelizable. It is also a concern of our service to be flexible, decoupled from specific processing nodes of the infrastructure and less structured as possible, attending different intelligent space applications and services. Since it can be easily found already installed in many different environments, a multi-camera system is used to overcome some difficulties traditionally faced by existing human detection methods that are based in only one camera. To validate our approach, we implement three different applications that are proof of concept (PoC) of many day-to-day real tasks. Two of these applications are related to robot navigation and demand the knowledge about the tridimensional localization of the humans present in the environment. With respect to time and detection performance requirements, our human detection service has proved to be suitable for interacting with the other services of our Intelligent Space, in order to successfully complete the tasks of each application. As an additional contribution, a feature extraction procedure based on the independent component analysis (ICA) theory is proposed as part of a detector and evaluated in public datasets. The pedestrian detection area is used as a playground to develop the human detector since it is the most mature research area of the human detection literature. The resulted detector is also used in the pipeline of the proposed human detection service, thus, being also applied in real-time applications in the intelligent space used as our testbed.

Lista de figuras

Lista de gráficos

Lista de tabelas

Lista de siglas e abreviaturas

ACF	*aggregate channel features*
BB	*bounding box*
ChnsFtrs	*channel features*
CNN	*convolutional neural network*
DT	*decision trees*
DCNN	*deep convolutional neural network*
desvPad	desvio padrão
FPPI	*false positive per image*
FN	falsos negativos
FP	falsos positivos
FCF	*filtered channel features*
FPS	*frames per second*
GPU	*graphics processing unit*
HOG	*histograms of oriented gradients*
ICA	*independent components analysis*
ICCF	*independent components channel features*
IoU	*intersection over union*
LDCF	*locally decorrelated channel features*
LogAvrMR	*log average miss rate*
MR	*miss rate*
PR	precisão
PdC	prova de conceito
RGB	*red-green-blue*
SOA	*service-oriented architecture*
SC	*sparse coding*
SVM	*support vector machine*
SNM	supressão de não máximo
3D	tridimensional
TP	*true positives*

Sumário

1 Introdução

1.1 Motivação

A detecção de seres humanos em imagens é uma importante área de pesquisa, devido à relevância do sensoriamento humano em diferentes situações do cotidiano (NGUYEN; LI; OGUNBONA, 2016). Nesse sentido, essa área tem sido adotada como tema central em uma variedade de estudos de caso.

A detecção de pedestres, por exemplo, tem sido estudada de forma recorrente na comunidade científica. Ao longo dos últimos anos, diversos trabalhos foram apresentados (DOLLÁR; BELONGIE; PERONA, 2010; BENENSON et al., 2015; ZHANG; BENENSON; SCHIELE, 2015; ZHANG et al., 2016a; ZHANG et al., 2016b; HU et al., 2017). Isso fez com que essa linha de pesquisa atingisse um alto grau de desenvolvimento, possuindo maior destaque quando comparada a outras tarefas relacionadas à detecção de seres humanos. O interesse particular pela detecção de pedestres se deve, principalmente, ao crescente desenvolvimento de pesquisas na área de veículos autônomos e sistemas avançados de assistência ao condutor (*advanced driver assistances systems*) (ADAS) (GERONIMO et al., 2010).

Apesar desse fato, a detecção de seres humanos em imagens é uma área que, certamente, não se limita a aplicações voltadas para veículos autônomos. Tarefas que envolvem câmeras de videomonitoramento também demandam desenvolvimento de pesquisas nessa área (VARGA; SZIRÁNYI, 2017). Nesse contexto, aplicações relacionadas a segurança público-privada (PATIL; TALELE, 2015), monitoramento de pessoas em multidões (ALAHI; RAMANATHAN; FEI-FEI, 2017) e detecção de eventos em estações de metrô (LIU et al., 2005) podem ser mencionadas. De uma forma mais abrangente que a de aplicações voltadas para videomonitoramento, encontra-se o desenvolvimento de soluções com foco em espaços inteligentes. Nessa área de estudo, o objetivo é, em geral, extrair informações do ambiente de forma a interagir com os indivíduos nele presentes (LEE; NORDSTEDT; HELAL, 2003). Dessa forma, a localização e o comportamento de seres humanos e objetos, presentes no espaço analisado, são de suma importância.

Um exemplo de aplicação voltada para espaços inteligentes, que envolve a detecção de seres humanos, é o desenvolvimento de robôs de serviço (PYO et al., 2015; GLAS et al., 2013; WANG et al., 2012). Entre outras funções, esses robôs podem ser utilizados para auxiliar seres humanos em ambientes como *shoppings* (GLAS. et al., 2015) e supermercados (MATSUHIRA et al., 2010). Nessas aplicações, o contexto no qual o ser humano está inserido é muito importante para que o robô possa prover o serviço correto, para a situação demandada. Nesses casos, a informação sobre a presença de pessoas no espaço de trabalho também é fundamental, pois, por meio dela, pode-se realizar o planejamento adequado da navegação dos robôs empregados (RIBEIRO et al., 2017).

Uma outro motivação, para a elaboração de soluções relacionadas à detecção de pessoas em ambientes inteligentes, é a construção de casas e salas de reunião automatizadas (HELAL et al., 2005; CHEN et al., 2004). Esses ambientes devem ser capazes de identificar as necessidades e interagir com os indivíduos neles presentes. Inseridos nesse mesmo contexto, encontram-se ainda aplicações voltadas para assistência a pessoas com necessidades específicas, como idosos (ALBAWENDI et al., 2015), e o desenvolvimento de sistemas inteligentes de vigilância (FREJLICHOWSKI et al., 2014).

A detecção de seres humanos é também importante em espaços inteligentes que possuam tarefas que demandem a interação homem-máquina, conforme verificado em (CHRUNGOO; MANIMARAN; RAVINDRAN, 2014; AKKALADEVI; HEINDL, 2015; PEREIRA; VASSALLO; SALLES, 2013). A extração do contexto sobre a cena e o reconhecimento de atividades são fundamentais para a execução dessas tarefas. Por fim, em uma escala superior, porém correlata a de espaços inteligentes, está o desenvolvimento de aplicações que envolvem detecção de seres humanos em cidades inteligentes (BOUTSIS; KALOGERAKI; GUNO, 2016).

Nesse sentido, a detecção de seres humanos em imagens, neste trabalho, está voltada para aplicações relacionadas a espaços inteligentes. Para esse fim, no entanto, serão utilizados conceitos e soluções relacionadas à detecção de pedestres, devido à maturidade dessa categoria da detecção de seres humanos na literatura.

Espaços inteligentes e a detecção de seres humanos

Na década de 90, Weiser sugeriu que a tecnologia do futuro estaria tão imersa na vida das pessoas que ela seria imperceptível (WEISER, 1991). Naquele tempo, espaços inteligentes não haviam ainda sido implementados, embora computadores pessoais estivessem já sendo produzidos e comercializados. Weiser afirmou que uma boa ferramenta deveria ser invisível e que, por isso, os computadores não seriam ainda a ferramenta ideal (WEISER, 1994). Nesse sentido, a computação deveria ser empregada de uma forma que o usuário não a notasse, tampouco necessitasse de conhecimento técnico sobre ela. Por meio dessas ideias, Weiser definiu o que é chamado de computação ubíqua (*ubiquitous computing*).

Estudos em computação ubíqua têm resultado em uma variedade de conceitos, entre eles estão os de espaços inteligentes (*intelligent spaces* e *smart spaces*) e ambientes inteligentes (*ambient intelligence*) (COEN, 1998; WRIGHT; STEVENTON, 2004; LEE; HASHIMOTO, 2002). Embora existam diferentes definições, todas possuem como fundamento básico a imersão da computação no ambiente. Sendo assim, tendo em vista que o objetivo deste trabalho não é discutir diferenças terminológicas no âmbito da computação ubíqua, o termo espaço inteligente será utilizado de forma geral.

Atualmente, espaços inteligentes tem atraído crescente interesse por parte da comunidade acadêmica. Nos trabalhos desenvolvidos, busca-se transferir para o ambiente inteligência e capacidade computacional, reduzindo a demanda por estações centrais com alto poder de processamento. Em geral, o objetivo final é extrair informações sobre o ambiente, executar todo

o processamento demandado e interagir provendo serviços para os usuários dessa plataforma (LEE; NORDSTEDT; HELAL, 2003).

Nesse sentido, é interessante que aplicações voltadas para espaços inteligentes possuam sistemas capazes de localizar pessoas em um determinado ambiente, de forma a extrair contextos por meio da observação do comportamento humano. Além disso, existe ainda a possibilidade de se oferecer serviços, por exemplo, por meio da interação homem-máquina, conforme observado em (GLAS. et al., 2015; GLAS et al., 2013; ZHOU et al., 2013; MRAZOVAC et al., 2012; BRSCIC, 2014).

De forma a extrair informação útil do ambiente e também detectar a presença humana, é comum o emprego de redes de sensores (CHEN; CHANG; CHEN, 2015; COOK et al., 2013; MRAZOVAC et al., 2012). Alguns trabalhos evitam o uso de câmeras devido a complexidade dos algoritmos envolvidos, problemas com oclusão e variações no brilho e cor das imagens. Além disso, a mudança na forma e aparência dos objetos, devido ao ângulo de visão, é um desafio. Nessas situações, o emprego de outros sensores, tais quais infravermelho, laser, de aceleração e movimento, torna-se uma alternativa (SURIE; PARTONIA; LINDGREN, 2013; MORIOKA; HASHIKAWA; TAKIGAWA, 2013; COOK et al., 2010).

No entanto, em situações em que se faz necessária uma interação direta com seres humanos, apenas a informação de presença pode não ser suficiente. Informações adicionais como localização e também reconhecimento de faces e gestos podem ser demandadas. Nesses casos, opta-se geralmente pelo emprego de câmeras, devido à maior riqueza das informações providas por esse tipo de sensor sobre o ambiente e os objetos nele presentes (LEE et al., 2012; ZABULIS et al., 2013).

Em se tratando especificamente da detecção de seres humanos em imagens, a maior parte dos métodos é fundamentada em modelos baseados em aparências. Essas técnicas empregam, em sua maioria, extratores de características associados a classificadores, que são treinados e avaliados em bases de dados públicas após dias ou horas de sintonia de hiperparâmetros. Conforme observado em (LI; YAO; WANG, 2012), em geral, esses tipos de detectores procuram por uma solução genérica[1]. Busca-se reduzir o erro de detecção o tanto quanto for possível, sem tornar o detector específico para uma determinada situação ou ambiente. Exceto em alguns poucos casos, apenas testes *off-line* em bases de dados públicas são conduzidos. Nos últimos anos, a maior parte da literatura tem se dedicado a intervenções na etapa de extração de características do fluxograma do processo de detecção, conforme observado em (BENENSON et al., 2015; CAO; PANG; LI, 2016).

No entanto, mesmo com o progresso da área de detecção de objetos, liderada pela família de soluções baseadas em redes neurais convolucionais (GIRSHICK et al., 2014), a tarefa de detectar pessoas e objetos em aplicações de tempo real, com qualidade, tal como exigido em espaços

[1] O termo "detector genérico", neste trabalho, será usado para se referir a métodos que são desenvolvidos sem qualquer restrição, *a priori*, no espaço de soluções das tarefas para as quais foram desenvolvidos. Dessa forma, não são levadas em conta no desenvolvimento desses detectores, por exemplo, informações acerca da geometria da cena.

inteligentes, é um problema ainda não resolvido. A utilização de detectores em aplicações práticas também carece de maior análise por parte dos trabalhos apresentados na literatura. Isso é mais evidente ainda para o caso em que uma rede de câmeras é empregada. Em geral, o processo de detecção em uma única câmera já demanda uma alta capacidade computacional, de forma que executar esse processo em uma rede integrada de câmeras torna-se algo ainda mais desafiador. Sendo assim, um dos problemas existentes, ao se aplicar a detecção de seres humanos em um espaço inteligente baseado em uma rede de câmeras, é desenvolver um detector de baixo custo computacional, que não sobrecarregue a rede, preservando assim recursos para outros serviços e aplicações.

Apesar dos problemas elencados, existem também vantagens em se implementar um sistema de detecção de seres humanos em um espaço inteligente. Se conceitos como computação em nuvem (*cloud computing*) e arquitetura orientada a serviços (*service-oriented architecture* - SOA) forem utilizados na concepção da plataforma desse espaço inteligente, a depender de seu projeto, o detector apresentará características como paralelismo, confiabilidade e alocação dinâmica de recursos. O detector de seres humanos poderá então se beneficiar de todas as vantagens dessa arquitetura. Nesses casos, a detecção de seres humanos pode ser fornecida como um serviço, que pode ser utilizado por diferentes aplicações e funcionar em diferentes plataformas, sendo ao mesmo tempo leve e flexível.

1.2 Proposta desta tese

Nesta tese de doutorado, um serviço de detecção de seres humanos leve, flexível e que funciona em um espaço inteligente é proposto. Esse serviço emprega uma arquitetura baseada em computação em nuvem e SOA, possui acoplamento mínimo à infraestrutura física e atende a diferentes aplicações de forma modular. A informação visual do ambiente é fornecida empregando apenas uma rede de câmeras. Além de prover detecção de seres humanos, o serviço agrega e transfere para as aplicações, que o utilizam, importantes características disponibilizadas pela plataforma do espaço inteligente e pela rede de câmeras. Diferentes aplicações que são provas de conceito (PdC) de tarefas do dia a dia, e que se baseiam na detecção de seres humanos, podem se beneficiar dessas características de forma a serem executadas de maneira mais rápida, flexível e confiável.

Em suma, os seguintes pontos são abordados na concepção do serviço de detecção de seres humanos proposto:

- A transferência, para as aplicações que utilizam o serviço, de propriedades como escalabilidade, paralelismo e confiabilidade, existentes no âmbito do espaço inteligente no qual é proposto;

- O acoplamento mínimo à infraestrutura física;

- A capacidade de localizar seres humanos em imagens advindas de uma rede de câmeras. Dessa forma, evitam-se problemas encontrados em abordagens com uma única câmera, que fazem detectores genéricos de seres humanos falharem em aplicações reais;

- O inter-relacionamento entre o serviço de detecção e a rede de câmeras, de forma a atender aos requisitos de tempo das aplicações. Isso permite que a execução em tempo real seja alcançada.

É importante mencionar que, de acordo com o apresentado em (ATIS, 2016), a distinção entre tempo real e aproximadamente tempo real é de certa forma nebulosa, devendo ser definida em cada contexto analisado. Neste trabalho, considera-se o termo tempo real como o alcance de taxas de processamento de dados, que não inserem atrasos perceptíveis nos requisitos de tempo das aplicações. Dessa forma, a experiência dos usuários, na interação com a infraestrutura do espaço inteligente, não deve ser afetada pelas técnicas de processamento de dados, desenvolvidas no âmbito desta tese. Um exemplo é o tempo de resposta do robô, em relação ao movimento dos seres humanos, nas aplicações desenvolvidas nesta tese de doutorado[2]. Idealmente, esse tempo de resposta não pode ser afetado significativamente pelo tempo de processamento do serviço de detecção de seres humanos.

O serviço de detecção, que é desenvolvido nesta tese de doutorado, possui como parte de seu fluxograma um detector genérico de seres humanos. Esse detector é proposto e avaliado no âmbito da detecção de pedestres e possui como diferencial, em relação a literatura, o procedimento de extração de características. Essa etapa tem sido considerada nos últimos anos como tendo papel de destaque para o sucesso do processo de detecção. O uso da teoria da análise de componentes independentes (*independent components analysis* - ICA) é avaliado para esse fim. A ICA, por ser uma técnica orientada aos dados, provê uma forma consistente de se realizar a extração de características. Finalmente, por meio das soluções propostas neste trabalho, pode-se fazer um paralelo entre detectores genéricos de seres humanos, avaliados geralmente apenas em situações controladas e testes *off-line*, e detectores aplicados a situações reais.

1.3 Contribuições

Consideram-se como contribuições principais desse trabalho:

- Um serviço de detecção de seres humanos flexível, que é adequado para ser utilizado por diferentes aplicações em um espaço inteligente baseado em uma rede de câmeras. A flexibilidade desse serviço é alcançada devido ao emprego de conceitos de computação em nuvem e SOA em seu desenvolvimento. Um acoplamento mínimo a infraestrutura física é obtido;

[2] Conforme será apresentado adiante, duas das aplicações desenvolvidas neste tese de doutorado, no Capítulo 3, empregam um robô.

- Um serviço de detecção desenvolvido para ser escalável, confiável e paralelizável, de forma a atender as demandas de tempo e capacidade computacional das aplicações. Todas essas características são suportadas pela arquitetura e infraestrutura do espaço inteligente e transferidas para as aplicações por meio do serviço proposto;

- Avaliação prática do serviço por meio da implementação de três diferentes aplicações PdC de tempo real, que são desenvolvidas de forma distribuída no espaço inteligente experimental. Duas dessas aplicações envolvem inclusive interação homem-máquina, no sentido que um robô deve navegar, por um ambiente, tendo conhecimento sobre a localização tridimensional dos indivíduos nele presentes.

Além disso, são consideradas como contribuições adicionais e características do sistema desenvolvido nesta tese:

- Análise do uso da teoria da ICA em um detector de seres humanos, no âmbito da detecção de pedestres;

- Localização tridimensional (3D) de indivíduos, no espaço inteligente, com erro médio menor que 40% do diâmetro médio da projeção da área do corpo humano no plano do chão. Essa característica é obtida principalmente pelo uso de uma rede de câmeras;

- Análise extensiva da acurácia do detector de seres humanos como um serviço de extração de informação de contexto para aplicações reais;

- Disponibilização de mais de seis mil imagens anotadas para comparação.

Por fim, a pesquisa desenvolvida nesta tese de doutorado resultou na publicação de dois artigos em congressos e um artigo em periódico qualificado:

- Modelo Estatístico para Filtragem de Exemplos Negativos na Detecção de Pedestres, apresentado no XXI Congresso Brasileiro de Automática – (ALMONFREY et al., 2016a);

- *Neural Cells Insights On Pedestrian Detection*, também apresentado no XXI Congresso Brasileiro de Automática – (ALMONFREY et al., 2016b);

- *A flexible human detection service suitable for Intelligent Spaces based on a multi-camera network*, publicado na revista *International Journal of Distributed Sensor Networks* – (ALMONFREY et al., 2018a).

1.4 Organização do trabalho

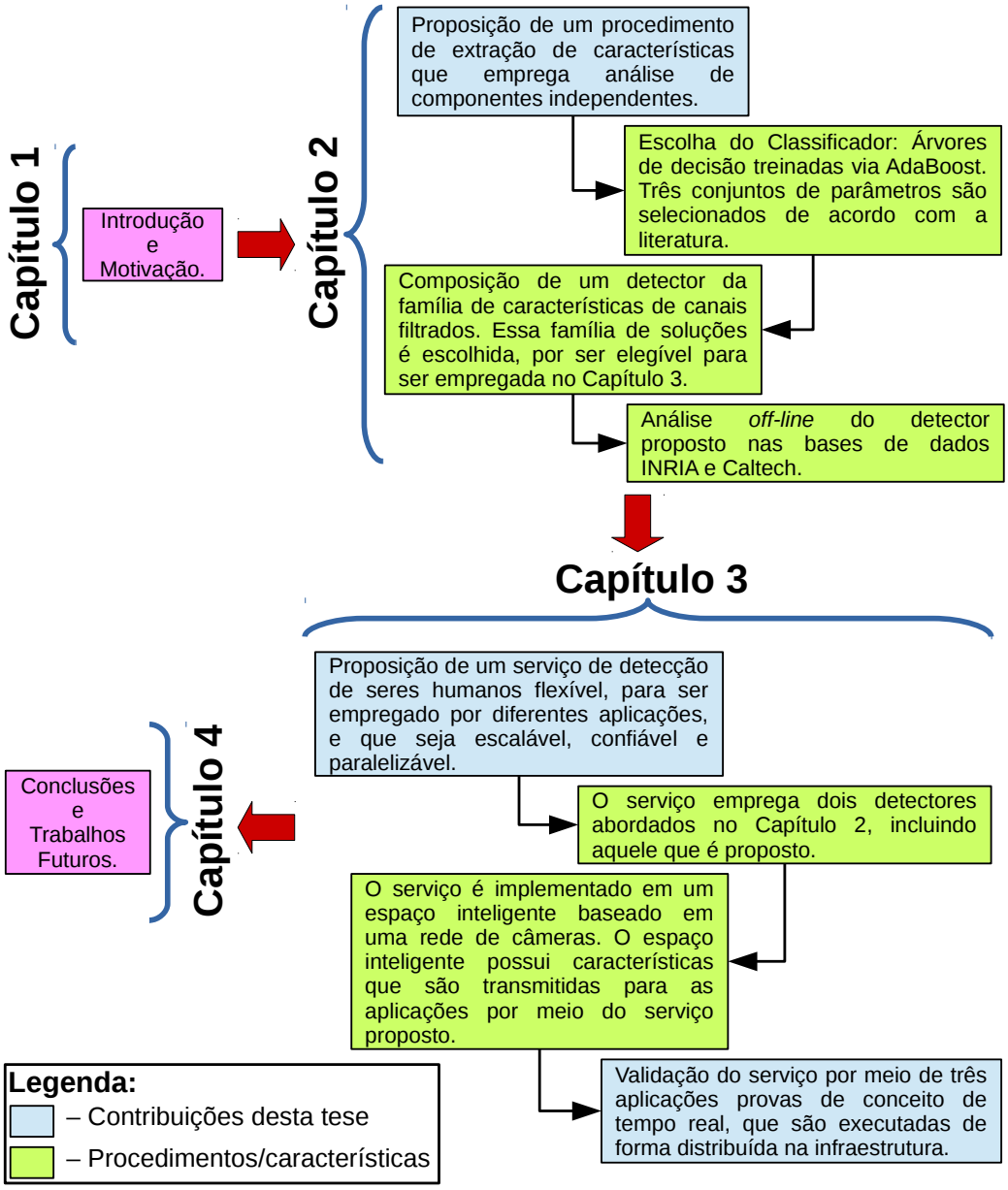
Este documento está organizado da seguinte maneira. No Capítulo 2, o detector de seres humanos proposto neste trabalho é analisado no âmbito da detecção de pedestres. Nessa ocasião,

aborda-se a forma como a ICA é empregada na tarefa de extração de características. Testes em um cenário *off-line* são conduzidos empregando bases de dados disponíveis na literatura.

No Capítulo 3, o problema de detecção de seres humanos é abordado no contexto de aplicações reais, em um cenário *on-line*. Um detector de seres humanos é então proposto em uma arquitetura baseada em serviços, com acoplamento mínimo a infraestrutura física e utilizado de forma modular por diferentes aplicações PdC. Uma análise sobre o desempenho de detectores genéricos de seres humanos, quando empregados em aplicações de tempo real, é realizada.

Por fim, no Capítulo 4, são apresentadas as conclusões deste trabalho, bem como algumas possibilidades de trabalhos futuros. A Figura 1 apresenta, de forma mais detalhada, a sequência na qual o trabalho desenvolvido nesta tese é apresentado. Essa figura também é utilizada para destacar algumas características, procedimentos e contribuições desta tese.

Figura 1 – Estrutura de apresentação do trabalho.



2 Detecção automática de seres humanos em imagens

Neste capítulo, uma proposta de solução para a tarefa de detecção de seres humanos em imagens é apresentada no contexto da detecção de pedestres. O detector proposto pertence a família de soluções conhecida como características de canais filtrados. Essa família possui métodos elegíveis para serem empregados no Capítulo 3, pois apresenta um bom compromisso entre qualidade e tempo de detecção. Além disso, métodos dessa família não dependem de *hardware* específico, como unidades de processamento gráfico. Ao longo deste capítulo, a teoria empregada no desenvolvimento do método e seu fluxo de operação são discutidos. Por fim, uma análise das potencialidades e limitações do detector proposto é realizada, por meio da comparação com trabalhos linha base de referência.

2.1 Trabalhos relacionados

A detecção de seres humanos em imagens é uma instância da detecção geral de objetos. Ambas possuem soluções enquadradas nas mesmas categorias. As técnicas de detecção de objetos em imagens podem ser, simplificadamente, categorizadas em: correspondência entre silhuetas e modelo baseado em aparência.

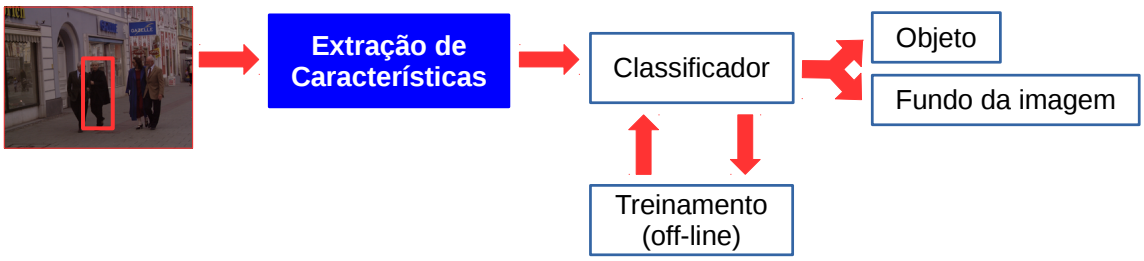
Na abordagem por correspondência entre silhuetas, em geral, emprega-se um modelo para representar a forma do objeto. Esse modelo pode ser estatístico (LIU; GUO; CHANG, 2014; ALMONFREY et al., 2016a) ou baseado na derivada ou gradiente de regiões da imagem, por exemplo, usando bordas (BROGGI et al., 2000). Em seguida, por meio de um processo de medida de similaridade, como a correlação, pode-se verificar quais partes da imagem correspondem ou não ao modelo, localizando-se assim o objeto na imagem. No entanto, os trabalhos mais bem sucedidos da literatura são fundamentados no modelo baseado em aparência, que também é empregado nesta tese de doutorado. Essa metodologia emprega características para representar as diferenças existentes entre regiões da imagem. Essas características são conhecidas também como descritores e são utilizadas para treinar classificadores, normalmente, de forma supervisionada.

A Figura 2 exibe de forma simplificada o fluxograma tradicional de um detector de objetos. Por meio de um conjunto de características obtidas para uma dada região da imagem, pode-se classificar essa região como sendo um objeto ou relativa ao fundo da cena. Juntos, os processos de extração de característica e classificação formam o núcleo da tarefa de detecção. Embora a qualidade do processo de detecção se deva em parte a um correto ajuste do classificador, conforme evidenciado em (PAISITKRIANGKRAI; SHEN; HENGEL, 2014; PAISITKRIANGKRAI; SHEN; HENGEL, 2016; ZHANG; BENENSON; SCHIELE, 2015; CAI; SABERIAN; VASCONCELOS, 2015; OHN-BAR; TRIVEDI, 2016), em (BENENSON et al., 2015) observa-se

que diferentes classificadores foram capazes de alcançar o estado da arte alterando, basicamente, os descritores empregados. Além disso, em (ZHANG et al., 2016a), observou-se que duas categorias diferentes de classificadores apresentaram desempenhos semelhantes quando empregando um mesmo conjunto de características.

Recentemente, alguns trabalhos como (ZHANG; BENENSON; SCHIELE, 2015; YANG et al., 2015; ZHANG et al., 2016a; CAO; PANG; LI, 2017; CAO; PANG; LI, 2016; HU et al., 2017) confirmaram a hipótese apresentada em (BENENSON et al., 2015), de que o projeto adequado de características possui papel de destaque no que diz respeito ao aumento da qualidade de detecção. Alguns resultados indicam ainda que é justamente a etapa de extração de características a principal responsável pelo sucesso das chamadas redes neurais convolucionais profundas (*deep convolutional neural networks* - DCNN) (YANG et al., 2015; ZHANG et al., 2016a; CAO; PANG; LI, 2017; HU et al., 2017).

Figura 2 – Fluxograma de um detector de objetos.



Detecção de pedestres em imagens

A relação com aplicações para carros autônomos, vídeo monitoramento e robótica torna a detecção de pedestres a classe da detecção de seres humanos mais abordada da literatura. Essa área de estudo é também empregada como *playground* para o desenvolvimento de novas técnicas de detecção de objetos (BENENSON et al., 2015). Devido a maturidade dessa área na literatura, ela é utilizada neste capítulo como estudo de caso para o problema de detecção de seres humanos em imagens.

Existem diferentes abordagens e modelos de soluções para o problema de detecção de pedestres na literatura. Neste trabalho, o escopo das soluções para o problema em questão se restringe àquelas que contribuem para a etapa de extração de características. Isso se deve a já mencionada importância dessa etapa no processo de detecção. Portanto, essa premissa é elemento norteador para a revisão bibliográfica a seguir, embora outros aspectos importantes dessa área de estudo sejam abordados em menor escala.

Métodos clássicos

Em (VIOLA; JONES; SNOW, 2003), apresentou-se um detector geral de objetos aplicado à tarefa de detecção de pedestres. Composto por um classificador com arquitetura em cascata

treinado via AdaBoost (FREUND; SCHAPIRE, 1997), o detector utiliza descritores gerados por padrões feitos a partir das funções de Haar. Esses padrões são arranjados espacialmente de maneira a modelar o corpo de um pedestre. Em outro trabalho, Dalal e Triggs (DALAL, 2006) propuseram o renomado e amplamente utilizado descritor baseado em histogramas de gradientes orientados (*histograms of oriented gradients* - HOG). Esses descritores foram durante muito tempo componentes base dos detectores de pedestres, embora nos últimos anos tenham perdido espaço para as DCNN.

Já em (WU; NEVATIA, 2007), empregaram-se características denominadas edgelets, de forma a modelar quatro partes do corpo de um pedestre em três diferentes poses. Por outro lado, em (FELZENSZWALB; MCALLESTER; RAMANAN, 2008), um esquema denominado modelo de partes deformadas (*deformable parts model* - DPM) foi apresentado. Nesse caso, o pedestre é representado por um modelo completo do corpo, de baixa resolução, juntamente com partes do corpo de alta resolução. As posições das partes são consideradas dinâmicas e desconhecidas. Como classificador, os autores utilizaram uma máquina de vetores suporte (*support vector machine* - SVM) específica para lidar com informações latentes.

Em (DOLLÁR et al., 2009a), o emprego de uma variedade de canais, derivados de componentes vermelho-verde-azul (*red-green-blue* - RGB) da imagem, foi avaliado na detecção de pedestres. O objetivo foi selecionar o melhor conjunto de canais a serem empregados como características. Desse estudo, 10 canais foram selecionados para a tarefa em questão: os canais de cor L, U e V; seis canais de HOG ($90^o, 60^o, 30^o, 0^o, -30^o, -60^o$) e um canal oriundo da magnitude do gradiente da imagem (|G|). Esses canais ficaram conhecidos na literatura como HOG+LUV e apresentaram os melhores resultados quando associados a árvores de decisão (*decision trees* - DT) treinadas via AdaBoost. O detector resultante dessa associação foi denominado características de canais (*channel features* - ChnsFtrs). Os canais HOG+LUV são também empregados na solução proposta neste trabalho e nos trabalhos linha base de referência: características de canais agregados (*aggregate channel features* - ACF) e características de canais localmente descorrelacionados (*locally decorrelated channel features* - LDCF).

Devido ao fato de obterem características por meio da filtragem (diversificação) de um conjunto base de canais (HOG+LUV), o método proposto neste trabalho e os detectores ChnsFtrs, ACF e LDCF são conhecidos como membros da família de soluções denominada características de canais filtrados (*filtered channel features* - FCF). Em (ZHANG; BENENSON; SCHIELE, 2015), essa família de soluções foi melhor definida e analisada. Seis diferentes tipos de filtros foram comparados quando aplicados aos canais HOG+LUV: InformedFilters, Checkerboards, RandomFilters, LDCF, LDCF8 e PCAForeground. As metodologias utilizadas para a obtenção desses filtros variam desde o emprego de padrões manualmente projetados com base no corpo humano à padrões escolhidos aleatoriamente. O conjunto de filtros mais bem sucedido para o problema abordado foi o Checkerboards, que consiste em amostras aleatórias de padrões de tabuleiro de xadrez de diferentes dimensões. É importante observar que, nesse caso, o conjunto selecionado de filtros não é dirigido aos dados, pois na metodologia de escolha não são

considerados os dados do problema.

Em (CAO; PANG; LI, 2016), dois conjuntos de filtros que avaliam regiões vizinhas e não vizinhas da imagem foram desenvolvidos no âmbito da detecção de pedestres. O primeiro leva em conta a aparência constante de três partes principais do corpo humano: cabeça, tronco e pernas. Por meio de filtros que avaliam regiões vizinhas e empregam operações diferenciais, pode-se destacar esse padrão do corpo humano em relação ao fundo da cena. O segundo conjunto de filtros opera em pares simétricos em relação ao eixo vertical da imagem. Dessa forma, captura-se a simetria apresentada pelo corpo humano, mas não presente na maior parte do fundo da cena.

Em (DOLLÁR et al., 2014), o bem sucedido, mas computacionalmente custoso processo de detecção multiescala[3], foi simplificado por meio de um esquema rápido de aproximação de canais de características. Ao invés de se realizar o cômputo das características para imagens amostradas em diferentes escalas, as características foram computadas para um conjunto de escalas de referência, sendo uma por oitava[4]. Em seguida, para as escalas intermediárias dentro de cada oitava, essas características foram aproximadas empregando uma lei de escalonamento exponencial. Dessa forma, múltiplas reamostragens da imagem foram evitadas. Esse esquema rápido de cômputo de características possibilitou ao detector ACF atingir taxas de detecção de 30 quadros por segundo (*frames per second* - FPS).

Abordagens baseadas em redes neurais convolucionais

Embora o emprego de redes neurais convolucionais (*convolutional neural networks* - CNN) na detecção de pedestres tenha sido questionado em (BENENSON et al., 2015), em (HOSANG et al., 2015) mostrou-se que essa abordagem pode ser competitiva com árvores de decisão treinadas via AdaBoost. No referido trabalho, algumas arquiteturas de CNN consagradas na literatura tiveram seus pesos ajustados para o problema de detecção de pedestres. Para isso, empregou-se um procedimento conhecido como transferência de aprendizado (*transfer learning*). De forma a melhorar a qualidade do processo de detecção, as CNN com pesos ajustados foram ainda associadas a classificadores SVM.

Mais recentemente, resultados apresentados na literatura indicam que a utilização de redes neurais convolucionais e arquiteturas profundas é crucial para se alcançar o estado da arte na detecção de pedestres. O emprego de CNN tem sido realizado, principalmente, na etapa de extração de características, provendo informação para um conjunto de árvores de decisão treinadas via AdaBoost (CAI; SABERIAN; VASCONCELOS, 2015; YANG et al., 2015; HU et al., 2017; CAO; PANG; LI, 2017; ZHANG et al., 2016a). Modelos baseados em partes têm também atraído atenção quando associados às DCNN (TIAN et al., 2015). O desenvolvimento de DCNN que sejam capazes de prover características invariantes à rotação tem também sido objeto de estudo nessa área (WENG et al., 2018).

[3] A detecção de pedestres deve ser realizada em múltiplas escalas de uma mesma imagem, haja vista que não se sabe o tamanho do pedestre na imagem *a priori*. Uma outra possibilidade, embora mais custosa computacionalmente, é treinar um classificador para cada escala possível do pedestre.

[4] Uma oitava é o intervalo entre uma escala e outra, que possui a metade ou o dobro do valor da primeira.

No entanto, o emprego de DCNN ainda depende de recursos computacionais específicos, como unidades de processamento gráfico (*graphics processing units* - GPU). Embora essa tecnologia tenha se tornado cada vez mais acessível, seu emprego em sistemas embarcados, carros autônomos ou espaços inteligentes nem sempre é a solução mais prática ou rentável. De certo, na busca pela solução de um problema, evitar a inserção de qualquer *hardware* específico em uma infraestrutura pré-existente, sempre que possível, é algo desejável. Conforme mencionado anteriormente, a utilização de soluções baseadas em CNN e DCNN é descartada nesse trabalho, devido a filosofia empregada na infraestrutura apresentada no Capítulo 3.

Emprego de dados adicionais

O uso de dados adicionais tem sido recorrente na literatura. Algumas técnicas que fazem uso desse tipo de informação adicional podem ser mencionadas: diferença entre quadros a partir de vídeos fracamente estabilizados (WALK et al., 2010), emprego de contexto (OUYANG; WANG, 2013) e fluxo óptico (WALK et al., 2010). Algumas abordagens empregam ainda outros tipos de sensores, por exemplo, câmeras térmicas (ZHAO et al., 2015) e sensores laser (PREMEBIDA et al., 2014). Soluções baseadas em DCNN têm também empregado imagens térmicas (LEE et al., 2015), além de técnicas envolvendo segmentação semântica como informação adicional (BRAZIL; YIN; LIU, 2017).

A utilização de técnicas de aumento de dados de treinamento, também conhecidas como *data augmentation* na literatura de aprendizado de máquina, tem se mostrado importante na tarefa de detecção de pedestres. Em (OHN-BAR; TRIVEDI, 2016), uma extensa análise, utilizando os detectores ACF e LDCF, foi realizada para uma nova configuração de parâmetros de um classificador, que é baseado em árvores de decisão treinadas via AdaBoost. Além disso, com a utilização de técnicas de aumento de dados de treinamento, os detectores ACF e LDCF se mostraram competitivos com os demais detectores da família de soluções de FCF. No referido trabalho, o detector LDCF se mostrou competitivo até mesmo com soluções baseadas em DCNN, quando utilizando um conjunto mais preciso de anotações de pedestres como dados de treinamento. Com essas contribuições, esses detectores foram chamados de ACF+ e LDCF+. Técnicas de aumento de dados de treinamento tem também sido úteis no treinamento de DCNN (ANGELOVA et al., 2015). Isso ocorre, porque nem sempre existem dados suficientes para se estimar os melhores parâmetros da rede.

Por fim, de forma a obter informações adicionais sobre a vasta literatura da detecção de pedestres, são sugeridos ao leitor os seguintes resumos (GERONIMO et al., 2010; DOLLÁR et al., 2012; BENENSON et al., 2015; ZHANG et al., 2016b).

2.2 Solução proposta

Conforme mencionado anteriormente, neste trabalho, a solução desenvolvida para o problema de detecção de seres humanos em imagens é analisada no contexto da detecção de pedestres.

Além disso, por ter papel de destaque nessa área de estudo, o projeto de características é o principal elemento considerado na construção dessa solução. Nesse sentido, dois detectores foram empregados como linha base de referência: ACF e LDCF.

Embora o ACF não seja o estado da arte da família de soluções de FCF, ele possui uma boa relação custo computacional *versus* qualidade de detecção. Isso o faz figurar entre as principais soluções da referida família, sendo um detector com taxa de processamento próxima a 30 FPS (DOLLÁR et al., 2014). Já o detector LDCF, por meio de sua versão apresentada em (OHN-BAR; TRIVEDI, 2016) (intitulada LDCF+), apresentou as melhores taxas de detecção para essa família de detectores.

Antes de apresentar a alteração proposta, neste trabalho, na etapa de extração de características, realiza-se a seguir uma apresentação sobre o funcionamento dos detectores linha base de referência.

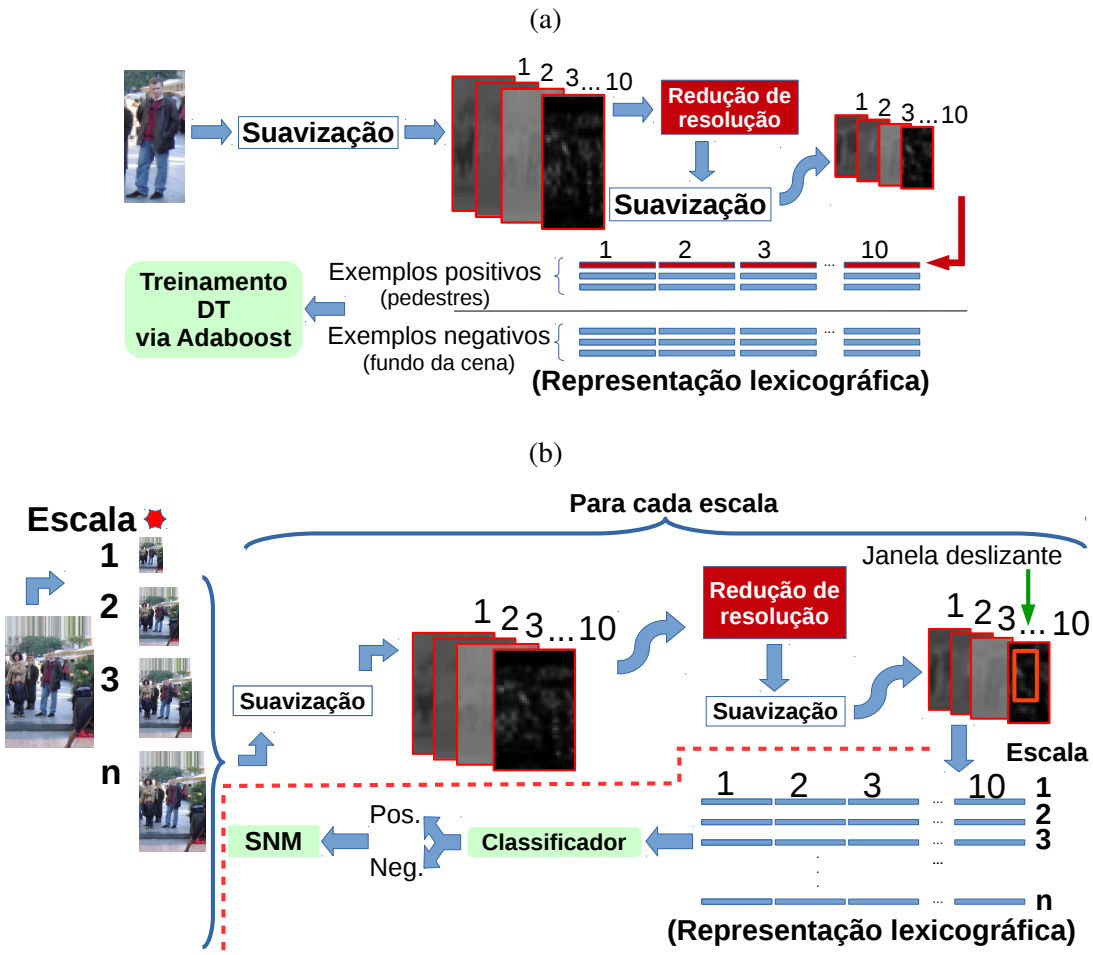
2.2.1 Detectores linha base de referência

Os fluxogramas do detector ACF, nas etapas de treinamento e teste, estão exibidos nas Figuras 3a e 3b, respectivamente. Na etapa de treinamento, os canais HOG+LUV são calculados após uma operação de suavização sobre a imagem. Em seguida, outra etapa de suavização é realizada, após uma operação de redução de resolução. Em (DOLLÁR et al., 2014), um fator de redução de resolução igual a quatro foi empregado. Por fim, os canais agregados são representados em sua forma lexicográfica (vetorizados) e utilizados junto ao processo de otimização AdaBoost durante o treinamento das árvores de decisão.

É importante salientar que cada píxel é tratado como uma característica única no processo de treinamento. A redução da resolução da imagem na Figura 3 é utilizada como uma forma de redução de dimensionalidade do vetor de características, ao custo de uma redução não significativa da qualidade de detecção (DOLLÁR et al., 2014). Observe ainda que, durante o treinamento, são necessárias amostras de imagens relativas a pedestres (exemplos positivos) e amostras de imagens onde nenhum pedestre está presente (exemplos negativos).

Como o detector não é treinado para realizar a detecção em múltiplas escalas de pedestres, os canais HOG+LUV devem ser computados em diferentes escalas na etapa de detecção. Desta forma, uma pirâmide de imagens é construída. Conforme mencionado anteriormente, o ACF emprega um processo rápido de cômputo de canais de características em diferentes escalas, por meio de uma lei de escalonamento exponencial (DOLLÁR; BELONGIE; PERONA, 2010). Esse processo evita reamostragens da imagem, tornando o processo de detecção mais rápido. A localização dessa etapa de cômputo rápido dos canais de características é representada no fluxograma da Figura 3b, por meio de um asterisco na cor vermelha. Após a classificação das janelas por meio de uma varredura empregando janelas deslizantes, um procedimento de supressão de não máximo (SNM) é executado de forma a remover detecções repetidas para um mesmo pedestre, incluindo em diferentes escalas.

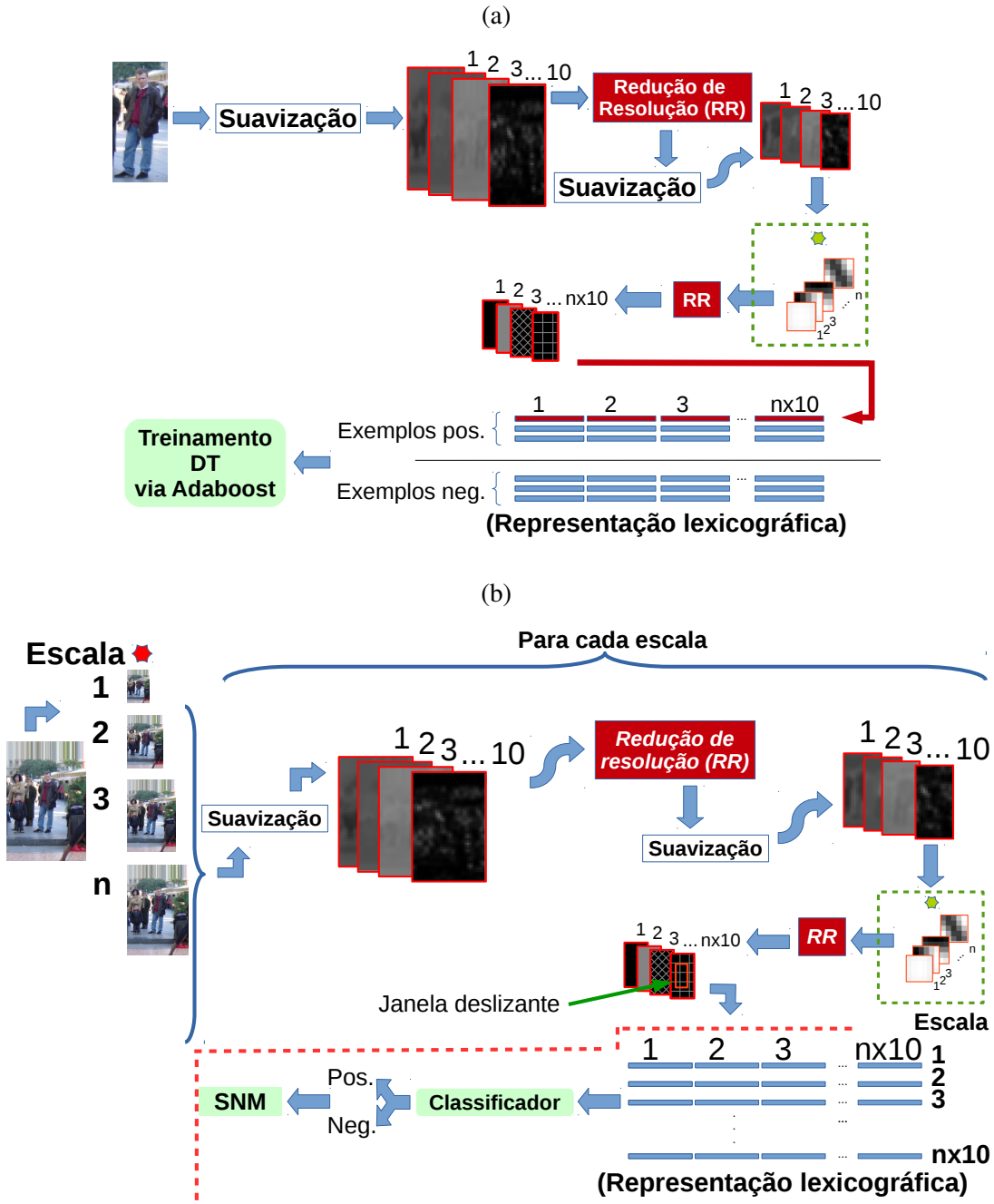
Figura 3 – Estágios de (a) treinamento e (b) teste do ACF.



O detector LDCF (NAM; DOLLÁR; HAN, 2014) emprega quatro filtros que proveem descorrelação local entre os píxeis da imagem, logo após a segunda etapa de suavização, nos estágios de treinamento e teste da Figura 3. A operação de redução da resolução, que antecede o processo de filtragem, é geralmente realizada por um fator igual a dois. Após a operação de descorrelação local, por meio da filtragem, outra operação de redução de resolução é realizada por um fator fixo igual a dois. Sendo assim, o fator total de redução de resolução é igual a quatro. As Figuras 4a e 4b exibem, respectivamente, as etapas de treinamento e detecção do LDCF. Repare, comparando as Figuras 3 e 4, que tanto na etapa de treinamento quanto na etapa de detecção, a única mudança em relação ao ACF é o processo de filtragem dos canais de características. Observe ainda, por meio da Figura 4, que um número arbitrário n de filtros pode ser empregado. Conforme mencionado anteriormente, para o LDCF, $\mathrm{n} = 4$. Contudo, nos experimentos realizados na Seção 2.3, diferentes valores serão avaliados para o literal n.

Os filtros de descorrelação utilizados no LDCF são autovetores da matriz de covariância dos dados, associados aos quatro maiores autovalores. O objetivo do LDCF é remover a correlação local entre os píxeis, aumentando a acurácia da classificação por parte de árvores de decisão com ramificações ortogonais. Essa categoria de árvores de decisão possui nós onde a classificação é baseada apenas em uma única característica. Nesse sentido, para esse tipo de árvore de

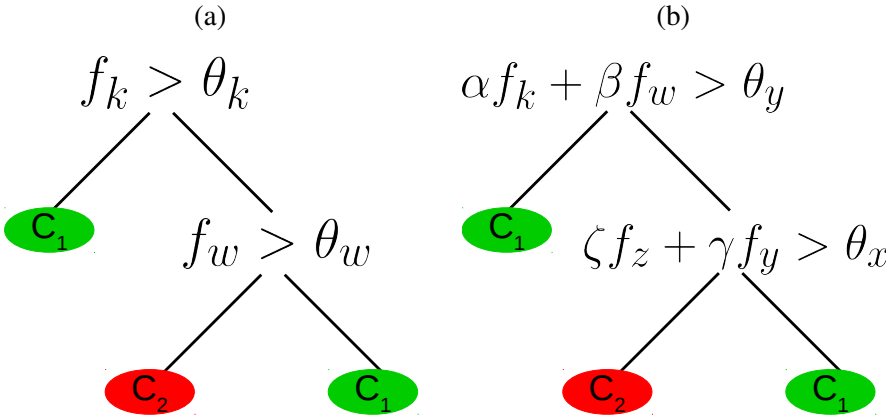
Figura 4 – Estágios de (a) treinamento e (b) teste do LDCF.



decisão, um maior desacoplamento entre as características é desejável. Árvores de decisão com ramificações ortogonais, usando dados descorrelacionados, superam em qualidade árvores com ramificações oblíquas, ao mesmo tempo que são computacionalmente mais eficientes (NAM; DOLLÁR; HAN, 2014). Árvores com ramificações oblíquas são empregadas quando os dados disponíveis são correlacionados. Esse tipo de árvore de decisão pode ser construído, por exemplo, para lidar com as relações lineares existentes entre as características. A Figura 5 exibe a diferença existente entre as árvores de decisão com ramificações ortogonais e oblíquas. Note que, diferentemente da árvore de decisão ortogonal, a oblíqua trata, no próprio nó, a relação

existente entre mais de uma característica. Por fim, no LDCF, um número final de 40 canais é obtido quando se aplicam quatro filtros de descorrelação a cada um dos 10 canais HOG+LUV. Esses canais são, em seguida, vetorizados e empregados junto a árvores de decisão treinadas via AdaBoost. Tanto o ACF quanto o LDCF possuem códigos-fontes disponíveis publicamente (DOLLÁR, 2016).

Figura 5 – Árvores de decisão com ramificações (a) ortogonais e (b) oblíquas. Repare que f_n é uma característica advinda de um vetor de características $\mathbf{f} = [f_1, f_2, ..., f_k, ..., f_n]$. Os literais C_1 e C_2 representam duas decisões possíveis em um problema de classificação binário, enquanto α, β, ζ, γ, θ_k, θ_w, θ_y e θ_x são constantes a serem estimadas no processo de aprendizado.



2.2.2 Uso do ICA na detecção de pedestres

Nesta seção, aborda-se a teoria envolvida na análise de componentes independentes (ICA). Além disso, a forma como esse conceito é empregado também é discutida. O detector LDCF será alterado para utilizar um processo de filtragem baseado na ICA, que é uma abordagem bioinspirada, baseada na forma como algumas células neurais representam seus dados. O objetivo é ter uma escolha mais fundamentada dos filtros utilizados, diferentemente do que foi feito em (ZHANG; BENENSON; SCHIELE, 2015), onde o filtro que resultou no detector mais bem sucedido (intitulado Checkerboards), foi escolhido sem qualquer consideração acerca da estrutura dos dados do problema.

No âmbito deste trabalho, a ICA é utilizada para remover a dependência local entre os píxeis da imagem, inspirada pelo detector LDCF, que emprega o conceito de análise de componente principal (*principal component analysis* - PCA), de forma a remover a correlação local entre esses píxeis. Usando ICA, espera-se representar dados multivariados em componentes tão independentes quanto possível, ao mesmo tempo que se aproveita da conexão existente entre os conceitos de ICA, codificação esparsa (*sparse coding* - SC) e teoria de células simples, quando aplicados a imagens naturais. Esse novo procedimento de filtragem irá simular, ainda que de forma simplificada, como o cérebro, mais especificamente o córtex visual, processa a informação. A forma como nosso sistema visual processa os dados é uma inspiração bastante intuitiva para

o projeto de técnicas de processamento de dados, como é o caso da geração de características direcionadas aos dados.

Imagens naturais (I) podem ser representadas por meio de uma base formada por um conjunto de m características, A_i, e componentes independentes, s_i,

$$I(x, y) = \sum_{i=1}^{m} A_i(x, y)s_i, \tag{1}$$

para todo x e y. Esta equação pode ser representada por uma transformação inversa

$$s_i = \sum_{x,y} W_i(x, y)I(x, y). \tag{2}$$

Em (HYVÄRINEN; HURRI; HOYER, 2009, Capítulos 7 e 8), é mencionado que, para imagens naturais, os coeficientes desconhecidos s_i são os mais independentes e esparsos quanto forem possíveis, quando os detectores de características W_i são obtidos por meio da minimização da entropia ou maximização da esparsidade, do conjunto de componentes lineares s_i. Desta forma, a Equação 2 se torna uma ICA e um SC da imagem. Quando obtidos de cenas naturais, esses modelos SC representam campos receptivos (*receptive fields*), similares aos das células simples presentes no córtex visual primário (V1) de macacos (ZHU; ROZELL, 2013; HYVÄRINEN; HURRI; HOYER, 2009, Capítulos 3 e 6).

Essa representação de mínima entropia da informação está relacionada ao consumo mínimo de energia do cérebro durante a atividade da região V1. No entanto, é importante mencionar que, conforme demonstrado em (ZYLBERBERG; DEWEESE, 2013), esses modelos SC não são a única forma de interpretar a computação sensorial do cérebro. Em alguns experimentos, animais denominados furões[5] demonstraram atividade neural densa (menos esparsa) na região V1, à medida que eles cresceram vendo cenas de vídeos naturais.

Os detectores de características W_i serão estimados, para cada um dos canais HOG+LUV, na etapa de treinamento. Em seguida, serão aplicados localmente a esses canais, nas etapas de treinamento e detecção, como máscaras de filtragem, em um processo similar ao LDCF (4). No entanto, ao invés de proverem descorrelação local, é esperado que esses filtros provejam independência local entre os píxeis dos canais. A ideia de independência é tida como tão importante ou mais que somente a descorrelação local, quando lidando com árvores de ramificações ortogonais.

Os filtros W_i são completamente dirigidos aos dados, sendo utilizados para sua estimativa o conjunto de dados separados para o treinamento do classificador. O detector gerado a partir do uso da ICA é denominado características de canais de componentes independentes (*independent components channel features* - ICCF) neste trabalho. Algumas máscaras geradas a partir dos detectores de características W_i podem ser vistas na Figura 6. Os filtros são gerados por meio do algoritmo FastICA, descrito em (HYVÄRINEN; HURRI; HOYER, 2009, Capítulo 18), cuja implementação encontra-se publicamente disponível (HYVÄRINEN, 2015).

Para redimensionar as máscaras para o tamanho requerido, a estratégia de reamostragem utilizando os vizinhos mais próximos é empregada. Essa abordagem apresentou resultados mais

[5] *Mustela putorius furo*.

consistentes que a baseada em interpolação. Como os filtros são esparsos, a interpolação parece atenuar a informação de magnitude. A Figura 7 contém os filtros relativos ao canal L da Figura 6, imediatamente antes do processo de reamostragem. Nas Figuras 6 e 7, a cor cinza representa valores iguais a zero, enquanto o branco representa valores positivos e o preto valores negativos. Repare, por meio das referidas figuras, que os filtros apresentam padrões que realizam o cômputo de gradientes locais, possuindo muitos valores próximos de zero.

Figura 6 – Filtros estimados usando ICA após a reamostragem. A cor cinza representa valores iguais a zero, enquanto o branco representa valores positivos e o preto valores negativos.

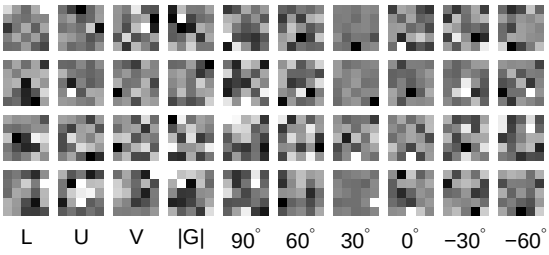
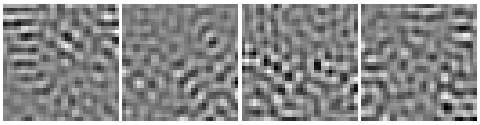


Figura 7 – Filtros do canal L antes da reamostragem. A cor cinza representa valores iguais a zero, enquanto o branco representa valores positivos e o preto valores negativos.



2.2.3 Classificador empregado

Dado que, neste trabalho, o projeto do detector possui como foco a etapa de extração de características, a escolha do classificador é fundamentada na literatura existente. Dentro do escopo da revisão bibliográfica da Seção 2.1, o classificador mais bem sucedido da família de soluções de FCF é composto por um conjunto de árvores de decisão, que são treinadas por meio do algoritmo AdaBoost. Esse classificador é também adotado nesta tese e uma breve explicação sobre a versão discreta do algoritmo AdaBoost, que é mais intuitiva, é apresentada a seguir.

Considere, primeiramente, um conjunto $S = \{x, y\}$, em que x representa um ou mais descritores e $y \in \{-1, +1\}$ um número que identifica a classe de um determinado objeto. É importante mencionar que x, em geral, representa um vetor de características com valores numéricos, que visam diferenciar os objetos a serem classificados. Como, no caso desta tese, o procedimento de classificação deve identificar quais regiões da imagem são pedestre ou não pedestre, os valores $\{-1, +1\}$ são tratados como {não pedestre, pedestre}. Considere ainda que

um conjunto de classificadores $f_t(x)$, para $t = 1...n$, compõem um classificador final $F(x)$, tal qual expresso na equação seguinte:

$$F(x) = \sum_{t=1}^{n} \alpha_t f_t(x), \tag{3}$$

sendo que $f_t(x)$ é o t-ésimo componente do modelo, com n componentes, e α_t uma constante. Cada um, dos n termos de $F(x)$, é uma árvore de decisão do conjunto total de árvores.

De forma a se estimar os parâmetros de $F(x)$, emprega-se o procedimento de otimização AdaBoost. Esse procedimento possui um algoritmo que ajusta um modelo de regressão logística aditivo $F(x)$, de acordo com a minimização da função critério exponencial $J(F)$ (FRIEDMAN; HASTIE; TIBSHIRANI, 2000), que é apresentada na equação seguinte:

$$J(F) = E[e^{-yF(x)}], \tag{4}$$

sendo que $E[w]$ é o valor esperado de w que, dependendo do contexto, pode ser a média amostral.

Pode-se demonstrar que a função critério $J(F)$ da Equação 4 possui valor mínimo (min) para:

$$F(x)_{min_{J(F)}} = \frac{1}{2} log \frac{P(y = 1|x)}{P(y = -1|x)}, \tag{5}$$

sendo que $P(y = c|x)$ é a probabilidade condicional de classe ($c \in \{-1, 1\}$) demandada em um problema de classificação. Logo,

$$P(y = 1|x) = \frac{1}{1 + e^{-2F(x)}}, \tag{6}$$

para $P(y = 1|x) + P(y = -1|x) = 1$. Repare que a Equação 6 difere do modelo logístico apenas por um fator igual a dois no termo exponencial. Desta forma, após o processo de otimização, por meio do algoritmo AdaBoost, tem-se $F(x) = \sum_{t=1}^{n} \alpha_t f_t(x) \approx F(x)_{min_{J(F)}}$. Com isso, pode-se empregar no processo de decisão a função sinal de um valor w, $sign(w)$, tal como apresentada na equação seguinte:

$$c = \begin{cases} 1, & \text{se } sign(F(x)) > 0; \\ -1, & \text{se } sign(F(x)) < 0. \end{cases} \tag{7}$$

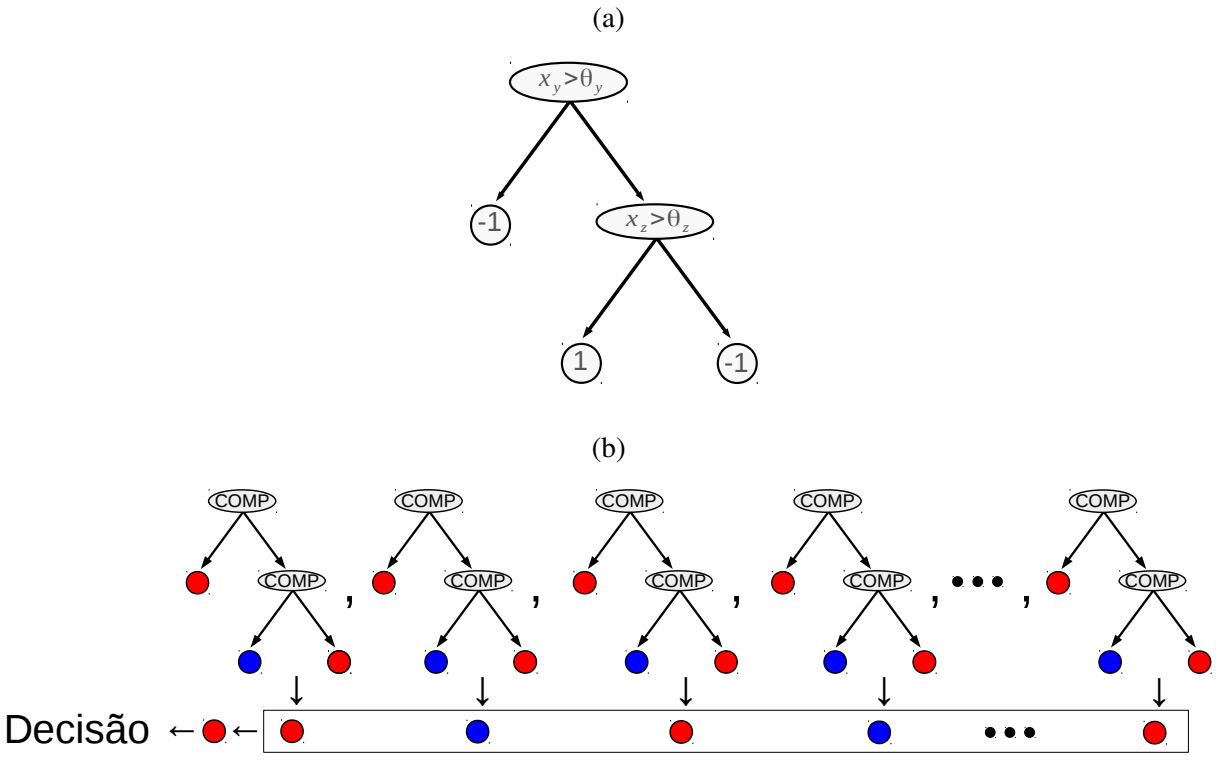
Para o problema binário discreto, é possível perceber que um valor positivo de $F(x)$, segundo a Equação 5, implica em $P(y = 1|x) > P(y = -1|x)$, enquanto o contrário é representado por um valor negativo de $F(x)$.

Nesta tese, os termos que compõem o modelo aditivo $F(x) = \sum_t \alpha_t f_t(x)$ são árvores binárias de decisão ortogonais. Conforme mencionado anteriormente, essas árvores são denominadas ortogonais, pois consideram valores relativos apenas a uma única característica em cada nó de decisão (Figura 8a). Esses tipos de árvores são, de maneira isolada, classificadores fracos com alto viés e baixa variância. A AdaBoost auxilia no processo de redução do viés na classificação, compondo assim um classificador final $F(x)$ forte. Além disso, em seu processo de otimização

iterativo, a AdaBoost possui uma esquema de ponderação, aumentando a importância dos exemplos erroneamente classificados (FRIEDMAN; HASTIE; TIBSHIRANI, 2000).

É importante ressaltar que a escolha da característica x_z de um dado nó, bem como seu limiar correspondente θ_z, faz parte do processo de treinamento da árvore. Cada árvore $f_t(x)$ possui seu processo individual de treinamento, com o treinamento global de $F(x)$ sendo conduzido pelo algoritmo AdaBoost. Cada termo da coleção também possui como saída apenas dois valores discretos, $f_t(x) \in \{-1, +1\}$. Desta forma, $F(x)$ atua como uma coleção de votos fornecidos por cada $f_t(x)$, conforme apresentado na Figura 8b. O valor de $F(x)$ é geralmente empregado como grau de confiabilidade relativo a uma decisão pela classe $c = 1$, ou seja, pedestre.

Figura 8 – Exemplo de (a) árvore de decisão e (b) de um conjunto de árvores de decisão. COMP se refere a uma comparação entre uma característica x_z e um parâmetro θ_z relativo a um dado nó da árvore. x_z e x_y são considerados como membros de um vetor de características $\mathbf{x} = [x_1, x_2, ..., x_n]$.



O critério de ramificação do processo de treinamento das árvores é baseado no erro de classificação, conforme descrito em (APPEL et al., 2013). Além disso, a busca pela característica ideal, a ser empregada em um determinado nó da árvore, pode ser feita em um subconjunto das características. Esse subconjunto é formado por meio da coleta aleatória de elementos do conjunto total de características empregadas. Esse procedimento evita a busca exaustiva na estimativa de parâmetros dos nós, reduzindo o tempo de treinamento.

Durante a etapa de predição, existe a possibilidade de se tomar a decisão sem avaliar todas as árvores do conjunto $F(x) = \sum_{t=1}^{m} \alpha_t f_t(x)$. Por exemplo, pode-se decidir pela classe $c = -1$ desde que o valor parcial de $F(x)$, $F(x)_p = \sum_{t=1}^{p} \alpha_t f_t(x)$ para $p < m$, esteja abaixo de um

dado limiar de rejeição (lim_{REJ}). Esse procedimento é chamado de limiar de rejeição rápido (DOLLÁR; APPEL; KIENZLE, 2012) e considera que uma vez que $F(x)_p < lim_{REJ}$, as demais árvores do conjunto não terão mais impacto sobre o resultado da decisão. Essa estratégia evita avaliações desnecessárias de árvores para exemplos negativos (maior parte das imagens) durante o processo de classificação, permitindo detecção em tempo real.

O algoritmo AdaBoost possui também um processo de otimização conhecido como versão real (RealBoost), no qual y não fica limitado ao conjunto $\{-1, 1\}$. Também é possível se obter variantes do processo de otimização AdaBoost que apresentam diferentes funções critério $J(F)$ (FRIEDMAN; HASTIE; TIBSHIRANI, 2000). Por fim, o procedimento de treinamento do classificador empregado, nesta tese, é o mesmo apresentado em (APPEL et al., 2013), cujo código-fonte encontra-se publicamente disponível em (DOLLÁR, 2016).

2.3 Experimentos

Nesta seção, uma análise da solução proposta na Seção 2.2 é conduzida no âmbito da detecção de pedestre. Essa análise é realizada por meio de um procedimento *off-line*, de forma controlada, por meio de bases de dados largamente empregadas na literatura.

2.3.1 Metodologia dos experimentos

O conjunto de experimentos realizado possui dois objetivos: (1) comparação do detector proposto com os detectores ACF e LDCF, membros da mesma família de soluções do ICCF. O objetivo dessa primeira análise é demonstrar que o ICCF é um detector competitivo com as principais soluções da família de FCF. Em termos de custo computacional e requisitos de *hardware*, essa família de soluções é adequada para ser utilizada no sistema desenvolvido no Capítulo 3. Taxas de FPS, que permitem execução em tempo real, são alcançadas sem a necessidade de emprego de GPU; (2) Por questões de completude, o desempenho do detector ICCF será comparado ao de outras técnicas da literatura, incluindo aquelas que empregam aprendizagem profunda (*deep learning*).

Bases de dados

Duas bases de dados são utilizadas ao longo dos experimentos:

- INRIA (DALAL, 2006): Um conjunto diverso de imagens estáticas cujo conjunto de teste é normalmente empregado como conjunto de validação (BENENSON et al., 2013; DOLLÁR et al., 2009a). A base de dados INRIA possui um conjunto de treino com 1237 exemplos positivos (pedestres) em 614 imagens, além de 1218 imagens onde nenhum pedestre está presente. O conjunto de testes contém 589 pedestres em 288 imagens e também 453 imagens sem qualquer pedestre. Neste trabalho, o conjunto de dados de teste

também será utilizado para validação de alguns parâmetros do processo de extração de características. A Figura 9a apresenta exemplos de imagens da base de dados INRIA.

- Caltech (DOLLÁR et al., 2009b): Conjunto de imagens extraídas a partir de um vídeo, obtido por meio de um veículo autônomo urbano. Essa base de dados possui imagens de pedestres em condições de detecção mais desafiadoras que a INRIA. Seis conjuntos (0-5) de dados com 192000 pedestres em 67000 imagens, e, também, 61000 imagens sem qualquer pedestre estão disponíveis para treinamento. Existem ainda cinco conjuntos de imagens (6-10) para teste, com 155000 pedestres em 65000 imagens, além de 56000 imagens sem qualquer pedestre. A Figura 9b apresenta exemplos de imagens da base de dados Caltech.

  Por ser disponibilizada na forma de um vídeo, é possível empregar diferentes intervalos de amostragem nessa base. Três configurações são utilizadas nos experimentos: Caltech-1Hz - configuração usual dessa base de dados, em que uma imagem é amostrada a cada 30 quadros. Essa é a única configuração empregada no conjunto de dados de teste; Caltech-7,5Hz - configuração intermediária, em que uma imagem é amostrada a cada quatro quadros; Caltech-10Hz - configuração mais densa que consiste na amostragem de uma imagem a cada três quadros. Além disso, um conjunto mais preciso de anotações da base de dados Caltech, disponibilizado em (ZHANG et al., 2016b), é empregado. Esse conjunto é denominado Caltech-10Hz$_{PREC}$.

  No âmbito dessa base de dados, são considerados na avaliação dos resultados pedestres enquadrados na condição denominada razoável, conforme definido em (DOLLÁR et al., 2012). Essa configuração apresenta pedestres com tamanho maior que 50 píxeis de altura, com parcial ou nenhuma oclusão.

Por fim, as duas bases de dados empregadas estão disponíveis publicamente em (CALTECH, 2009). Nesse repositório de dados, diferentes métodos da literatura são elencados de acordo com a acurácia.

Métricas

A métrica empregada para realizar a comparação entre os detectores é o logaritmo da média da taxa de perda (*log average miss rate* - LogAvrMR), obtida a partir da curva característica de operação do receptor (*receiving operating characteristics* - ROC) em que se tem a taxa de perda (*miss rate* - MR) *versus* falsos positivos por imagem (*false positive per image* - FPPI). A MR é a fração total de seres humanos não identificados e FPPI é o número de detecções que são falsos positivos dividido pelo número de imagens do experimento avaliado. A LogAvrMR é obtida a partir da média de nove pontos entre 10^{-2} e 10^{0} FPPI na curva ROC.

De forma a ser considerada uma detecção correta ou positiva, uma janela de detecção (*bounding box detection* - BB_{dt}) deve corresponder a alguma anotação de pedestre da base de dados

Figura 9 – Exemplos de imagens das bases de dados (a) INRIA e (b) Caltech.

(a)

(b)

(*bounding box ground truth* - BB_{gt}). Em caso de não haver qualquer correspondência, a janela de detecção é considerada uma detecção falso positiva. É importante mencionar que múltiplas correspondências a uma mesma BB_{gt} não são permitidas. Como medida de equivalência entre uma detecção e uma anotação, emprega-se a área de intercessão sobre união (*intersection over union* - IoU). Considera-se uma correspondência quando existe uma IoU maior que um limiar, lim_{IOU}, conforme representado na equação seguinte:

$$correspondência(BB_{dt}, BB_{gt}) = \frac{área(BB_{dt} \cap BB_{gt})}{área(BB_{dt} \cup BB_{gt})} > lim_{IOU}. \tag{8}$$

Essa equação é também conhecida como medida de Pascal e um limiar $lim_{IOU} = 0,5$ é empregado neste trabalho. É importante ressaltar que a metodologia de avaliação apresentada está em consonância com a definida em (DOLLÁR et al., 2009b), que é adotada como um padrão na área da detecção de pedestres.

A menos que se mencione o contrário, todos os experimentos são executados 10 vezes para diferentes sementes aleatórias (*random seeds*). Como resultados, são apresentados a média da LogAvrMR e o desvio padrão (desvPad) relativos às 10 rodadas de experimentos. Essa é uma forma mais robusta e justa de se realizar os experimentos. Uma quantidade significativa de trabalhos da literatura consideram apenas o melhor resultado como critério final de comparação entre diferentes detectores.

É importante mencionar que, em alguns casos, o melhor resultado pode ser significativamente

melhor que o valor médio, não podendo, portanto, ser considerado sozinho na análise comparativa. Essa metodologia experimental é de grande importância principalmente quando se trabalha com árvores de decisão. Esse classificador pode apresentar uma aleatoriedade na estimativa dos parâmetros de um nó, conforme apresentado na Seção 2.2.3.

Escolha dos hiperparâmetros e treinamento do classificador

Como o diferencial do detector proposto neste trabalho encontra-se na etapa de extração de características, busca-se, ao longo dos experimentos, restringir o procedimento de avaliação a essa etapa. Nesse sentido, o classificador escolhido possui hiperparâmetros selecionados de acordo com os trabalhos linha base de comparação. Três modelos são empregados e seus principais parâmetros são apresentados na Tabela 1. A versão real da AdaBoost é empregada no treinamento do classificador em todos os experimentos. Os demais parâmetros empregados no classificador são os mesmos utilizados no detector LDCF, cujo código-fonte encontra-se disponível em (DOLLÁR, 2016).

Tabela 1 – Parâmetros empregados nos modelos baseados em árvores de decisão.

Linha	Parâmetros	Modelo 1	Modelo 2	Modelo 3
1	Prof. Máx. Árvores	3	5	5
2	N.° Btrp./Est.	3/4	3/4	3/4
3	Fração Características Est.	1/16	1/16	1/16 e 1
4	N.° Árvores	2048	4096	4096
5	Dimensões Pedestre	128×64 (píxeis)	64×32	128×64
6	N.° Neg. Est./Máx.	5000/10000	25000/50000	100000/200000
7	N.° Árvores Est.	32, 128, 512 e 2048	64, 256, 1024 e 4096	64, 256, 1024 e 4096
8	Base de Dados	INRIA	Caltech-7,5Hz	Caltech-7,5Hz

* Abreviações: Neg – negativos, Btrp – *bootstrapping*, Est – estágio, Prof – profundidade e Máx – máximo(a).

De forma a entender melhor os parâmetros apresentados na Tabela 1, considere o modelo aditivo discutido na Seção 2.2.3: $F(x) = \sum_{t=1}^{m} \alpha_t f_t(x)$. O processo de treinamento das árvores $f_t(x)$, para $t = 1...m$, é realizado em N estágios. $N - 1$ estágios são utilizados como rodadas de *bootstrapping*, conforme representado na tabela pelos itens relativos à linha 2 (N.° Btrp./Est.). Nas rodadas de *bootstrapping*, utilizam-se modelos parciais do classificador, de forma a se obter exemplos negativos classificados como positivos. Esse procedimento é também conhecido como mineração de exemplos negativos difíceis (*hard negative mining*).

O modelo $F(x)$ possui um número m máximo de árvores (N.° Árvores - linha 4, Tabela 1). No entanto, ao longo dos estágios de treinamento do algoritmo AdaBoost, um número progressivo de árvores é empregado (N.º Árvores Est. - linha 7). O classificador final é o obtido no último estágio. Em cada estágio de treinamento, um novo número de negativos é adicionado ao número total de exemplos negativos acumulados (N.° Neg. Est./Máx. - linha 6). Esses exemplos adicionais são obtidos por meio do já mencionado procedimento de mineração de exemplos negativos difíceis. Quando o número de negativos acumulados supera o máximo possível, uma amostragem aleatória é realizada sobre esse conjunto, de forma a respeitar o limite máximo. Para o Modelo

1, por exemplo, na linha correspondente a N.° Neg. Est./Máx., 5000 amostras são adicionadas aos exemplos negativos até que o máximo de 10000 seja atingido. Após esse máximo, uma coleta aleatória é realizada. Essa adição e atualização dos exemplos negativos, no treinamento do classificador, torna o processo de rejeição de partes da cena, relacionadas ao fundo da imagem, mais eficaz.

Cada árvore de decisão binária $f_t(x)$ do modelo possui uma profundidade máxima especificada (Prof. Máx. Árvores - linha 1, Tabela 1). Conforme mencionado na Seção 2.2.3, de forma a reduzir o tempo de treinamento de cada nó da árvore, a busca exaustiva pela melhor característica, a ser empregada em cada nó, se dá em um subconjunto das características existentes. Esse subconjunto é formado por uma fração (Fração Características Est. - linha 3) amostrada aleatoriamente do conjunto total de características disponíveis. O Modelo 3 apresenta uma particularidade no que tange esse parâmetro, em relação aos Modelos 1 e 2. No último estágio de treinamento, as primeiras 512 árvores são treinadas por meio da busca exaustiva em todo o conjunto de características. Por isso, o parâmetro correspondente à linha (Fração Características Est.) é também igual a 1 para o Modelo 3. Em (OHN-BAR; TRIVEDI, 2016), a justificativa para essa alteração foi reduzir a influência da aleatoriedade no processo de treinamento.

O emprego de técnicas de aumento de dados de treinamento pode ser crucial para o desempenho do modelo estimado. Os modelos da Tabela 1 utilizam aumento de dados nos exemplos positivos. O Modelo 1 emprega espelhamento e translações de 0, -1 e 1 píxel na vertical e horizontal, de forma a gerar variações dos exemplos positivos disponíveis. O Modelo 2 emprega apenas espelhamento, enquanto o Modelo 3 emprega variações com escala aumentada em 10% na vertical, horizontal e em ambas as direções ao mesmo tempo.

Os Modelos 1 e 2 foram obtidos da análise realizada em (NAM; DOLLÁR; HAN, 2014), onde o detector LDCF foi proposto e comparado ao detector ACF. Esses conjuntos de parâmetros foram levantados para a base de dados INRIA (Modelo 1) e Caltech (Modelo 2). O Modelo 3 foi empregado em (OHN-BAR; TRIVEDI, 2016), junto à técnica de aumento de dados supracitada. Essa associação foi utilizada em variações dos detectores ACF e LDCF, que foram intituladas ACF+ e LDCF+, respectivamente. Esses detectores demonstraram ser competitivos com o estado da arte da família de FCF, sendo que o LDCF+ mostrou-se competitivo inclusive com métodos baseados em aprendizagem profunda, ao empregar o conjunto de anotações mais precisas Caltech-10Hz$_{PREC}$.

O Modelo 3 apresenta também, conforme sugerido em (OHN-BAR; TRIVEDI, 2016), algumas alterações relacionadas à etapa de extração de características. A primeira delas é que na pirâmide de imagens, formada no processo de detecção (Figuras 3b e 4b), considera-se também uma versão ampliada da imagem original, por um fator igual a dois, além de suas escalas intermediárias. Isso é feito para lidar com pedestres de menor resolução na imagem. Os Modelos 1 e 2 consideram apenas versões reduzidas da imagem original. Além disso, o processo de redução de resolução das Figuras 4a e 4b, são realizados de uma só vez, por um fator igual a quatro, logo após o processo de cômputo dos canais. Dessa forma, a etapa de redução de resolução, presente

após o procedimento de filtragem no ICCF e LDCF, é removida. O processo de aproximação dos canais de características, por meio da lei de aproximação exponencial, também é removido. Sendo assim, em todas as escalas, os canais de características são computados a partir de imagens reamostradas.

Por fim, a utilização de diferentes configurações de parâmetros dos classificadores permite uma comparação ampla dos diferentes processos de extração de características. Assim, pode-se analisar essa etapa para classificadores com diferentes capacidades de aprendizado.

2.3.2 Resultados experimentais

Base de dados INRIA

Modelo 1

Neste experimento, a análise do uso da teoria da ICA para estimar os filtros a serem aplicados aos canais HOG+LUV é realizada. O conjunto de teste da base de dados INRIA é utilizado como conjunto de validação para alguns parâmetros do processo de extração de características. Nesse ponto, uma comparação com o detector LDCF também é conduzida. O Modelo 1 da Tabela 1 é empregado. A quantidade de filtros e o tamanho das máscaras, bem como a extração de características utilizando subconjuntos dos canais HOG+LUV, são avaliados. A melhor configuração de parâmetros é então utilizada na base de dados Caltech.

No Gráfico 1a, o desempenho dos filtros ICCF são comparados ao do LDCF, quando varia-se a quantidade e o tamanho dos filtros. Como pode ser visto a partir do Gráfico 1a, para uma máscara 5×5, o melhor desempenho médio do ICCF fica próximo ao melhor resultado do LDCF, quando são empregados 18 e 25 filtros. É importante mencionar que, para uma máscara 5×5, o número máximo de filtros que o LDCF pode estimar é 25, de forma que, para se ter mais filtros, o tamanho da máscara precisa ser aumentado[6]. Isso é um problema pois, como é apresentado adiante, o aumento no tamanho da máscara corresponde a uma redução na qualidade da detecção. Logo, para o LDCF, aumentar o número de filtros para gerar mais características aparenta ser um problema, enquanto o mesmo não pode ser dito para o ICCF.

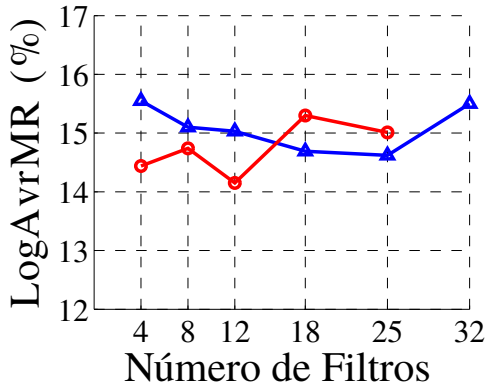
No Gráfico 1a, o ICCF tem um ligeiro aumento de desempenho médio ao se variar o número de filtros de quatro para 25, ao contrário do LDCF. Esse aumento no desempenho médio do ICCF, ao se utilizar mais filtros, sugere que a qualidade da detecção pode ser melhorada por meio de um aumento na capacidade do modelo $F(x)$[7]. Essa suposição se deve ao fato de mais características estarem disponíveis para serem selecionadas pela AdaBoost. Nesse caso, os dados empregados no treinamento do detector devem também ser aumentados, de forma a preencher esse maior modelo. Esta suspeita é averiguada para a base de dados Caltech, que permite diferentes intervalos de

[6] Devido a implementação da matriz de covariância dos dados do detector LDCF em (NAM; DOLLÁR; HAN, 2014), para uma máscara R×R, o número máximo de filtros que podem ser estimados é R^2.

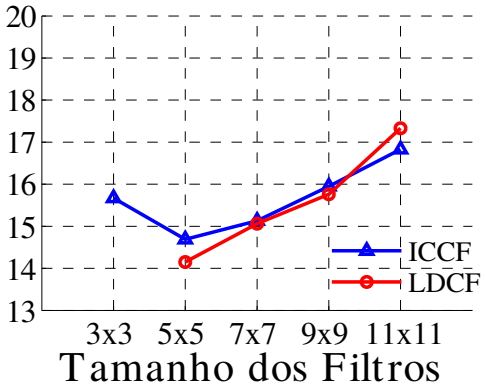
[7] O aumento da capacidade do modelo se refere, por exemplo, ao aumento da quantidade e da profundidade das árvores empregadas.

Gráfico 1 – Análise da variação (a) da quantidade e (b) dimensões dos filtros dos detectores LDCF e ICCF (INRIA).

(a) Empregando uma máscara 5×5.



(b) ICCF - 18 filtros e LDCF - 12 filtros.



amostragem. É importante notar também que, para 32 filtros, o desempenho médio do ICCF cai ligeiramente. Isso pode ser causado pelo tamanho limitado do Modelo 1 ou pela quantidade de dados disponíveis para treinar o detector. Neste caso, um número maior de características é gerado, sendo talvez necessário utilizar mais dados que os disponíveis na base de dados INRIA no treinamento.

Outra observação, obtida por meio do Gráfico 1b, é que o desempenho médio da detecção de ambos os métodos diminui à medida que as dimensões das máscaras aumentam. De fato, isso também é uma observação presente em (NAM; DOLLÁR; HAN, 2014), (ZHANG; BENENSON; SCHIELE, 2015) e (YANG et al., 2015). De acordo com esse último, isso parece estar relacionado à perda do foco em informações e especificidades locais como bordas e texturas, à medida que as dimensões aumentam. De forma a analisar o desempenho quando se aumentam as dimensões das máscaras, o número de filtros empregados no ICCF foi ajustado para 18, enquanto no LDCF foi ajustado para 12. Isso se justifica pois, para uma máscara 5×5, essas quantidades de filtros resultaram nos melhores desempenhos para os referidos métodos. Para uma máscara 3×3, o número máximo possível de filtros LDCF é nove e, então, não são apresentados resultados para essa configuração. Por simplicidade, a Tabela 2 apresenta a LogAvrMR e o desvPad do ICCF e LDCF para as diferentes quantidades e tamanhos de filtros, conforme o Gráfico 1. Para cada um dos métodos, a menor média da LogAvrMR é marcada em negrito nas Tabelas 2a e 2b.

Tabela 2 – Análise da variação de parâmetros dos detectores LDCF e ICCF (INRIA).

(a) Empregando uma máscara 5×5.

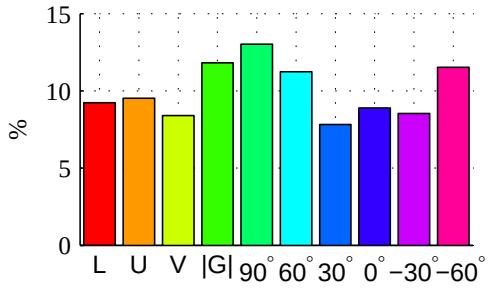
Número de filtros	LogAvrMR ± desvPad (%) ICCF	LDCF
4	15,55 ± 0,88	14,44 ± 0,71
8	15,10 ± 0,65	14,74 ± 0,59
12	15,03 ± 0,80	**14,15 ± 0,65**
18	14,69 ± 0,90	15,30 ± 0,82
25	**14,62 ± 1,01**	15,01 ± 0,87
32	15,50 ± 0,95	–

(b) ICCF - 18 filtros e LDCF - 12 filtros.

Tamanho dos filtros	LogAvrMR ± desvPad (%) ICCF	LDCF
3×3	15,67 ± 0,59	–
5×5	**14,69 ± 0,90**	**14,15 ± 0,65**
7×7	15,13 ± 0,79	15,06 ± 0,55
9×9	15,95 ± 1,18	15,76 ± 1,11
11×11	16,83 ± 1,02	17,33 ± 0,99

A melhor configuração observada para o ICCF é obtida quando se utiliza uma máscara de tamanho 5×5 e 25 filtros. No entanto, como o resultado para 18 filtros é muito próximo ao que se obtém empregando 25 filtros, caso não se mencione o contrário, serão utilizados 18 filtros nos testes a seguir. Essa escolha se deve ao fato do aumento no número de filtros implicar em um aumento nos tempos de treinamento e detecção. Sendo assim, para uma máscara 5×5 e 18 filtros, o Gráfico 2 exibe o percentual de características originárias dos canais HOG+LUV, que foram selecionadas pela AdaBoost após o processo de filtragem realizado no detector ICCF. Repare que a utilização das características é relativamente bem distribuída no histograma apresentado, indicando uma baixa redundância entre as informações geradas a partir de cada canal.

Gráfico 2 – Percentual de características, relacionadas a cada um dos canais, selecionadas pela AdaBoost (INRIA).



A Tabela 3 mostra o efeito do ICCF quando se usam diferentes configurações de canais e também quando os canais gerados por esse detector são empregados em conjunto com os canais gerados pelo LDCF. Na primeira linha, a coluna 2, N^o de canais, está representada como 18×3 (LUV)+7 (|G| HOG). Isso significa que os 18 filtros relativos ao ICCF foram aplicados apenas aos três canais LUV, enquanto os canais de magnitude do gradiente |G| e de HOG ($90^o, 60^o, 30^o, 0^o, -30^o, -60^o$) (sete) não passaram pelo processo de filtragem. Sendo assim, para esse caso, 61 canais estarão disponíveis. Para as demais linhas, o mesmo raciocínio deve ser utilizado. Repare que o resultado da aplicação do ICCF somente aos quatro canais mais utilizados do Gráfico 2 também é apresentado na terceira linha (|G| 90°60°-60°). Fica claro que a aplicação da ICA aos canais relacionados ao gradiente (sete canais) é mais importante que apenas aos canais LUV (três canais). No entanto, o melhor resultado médio é obtido quando todos os canais são utilizados, mostrando a complementariedade entre eles.

Tabela 3 – Filtros ICCF empregados em diferente configurações de canais (INRIA).

Configuração	N° de canais	LogAvrMR ± desvPad (%)
LUV	18×3 (LUV)+7 (|G| HOG)	17,97 ± 0,87
|G| HOG	3+18×7	15,74 ± 1,5
|G| 90°60°-60°	18×4+6	16,92 ± 0,67
HOG+LUV	18×10	14,69 ± 0,90
ICCF+LDCF	18×10+4×10	**14,13 ± 0,76**
ICCF+LDCF	18×10+12×10	14,56 ± 0,76

É também importante mencionar que as duas últimas linhas da Tabela 3 apresentam os resultados do ICCF (18 filtros) quando empregado em conjunto com o LDCF (quatro ou 12

filtros). Os vetores de características de cada uma dessas duas abordagens são concatenados, formando um único conjunto de características. Esse procedimento pode ser entendido como uma fusão entre características distintas. Pode-se observar alguma complementariedade entre o ICCF e LDCF (quatro filtros) - (22 no total) porque a média da LogAvrMR diminui de 14,44% (considerando apenas o LDCF - Tabela 2a) para 14,13% (ICCF+LDCF - Tabela 3). No entanto, a mesma situação não ocorre para o caso do ICCF e LDCF (12 filtros) - (30 filtros no total), cuja média da LogAvrMR aumenta de 14,15% (somente LDCF - Tabela 2a) para 14,56% (ICCF+LDCF - Tabela 3).

De forma a facilitar a análise comparativa, a Tabela 4 apresenta os melhores resultados médios dos detectores LDCF e ICCF. O melhor resultado médio para o ACF também é apresentado. Repare que o LDCF é apresentado em suas versões com quatro (LDCF-4) e 12 (LDCF-12) filtros. Embora a versão com 12 filtros apresente um resultado ligeiramente superior, é a versão com quatro filtros que foi originalmente proposta em (NAM; DOLLÁR; HAN, 2014). Essa versão original foi retreinada neste trabalho. Por meio dessa tabela, percebe-se que tanto o LDCF quanto o ICCF possuem desempenho superior ao ACF. Em termos de desempenho médio, a versão LDCF-4 é ligeiramente superior ao ICCF, embora essa diferença esteja contida dentro do desvio padrão das rodadas do experimento. Por meio dos resultados da Tabela 4, pode-se afirmar que o ICCF é um detector competitivo com o LDCF em termos de LogAvrMR.

Tabela 4 – Desempenhos médios do ACF, LDCF e ICCF (INRIA).

Detector	ACF	LDCF-4	LDCF-12	ICCF
LogAvrMR $\pm$ desvPad (%)	17,96 $\pm$ 1,19	14,44 $\pm$ 0,71	14,15 $\pm$ 0,65	14,69 $\pm$ 0,90

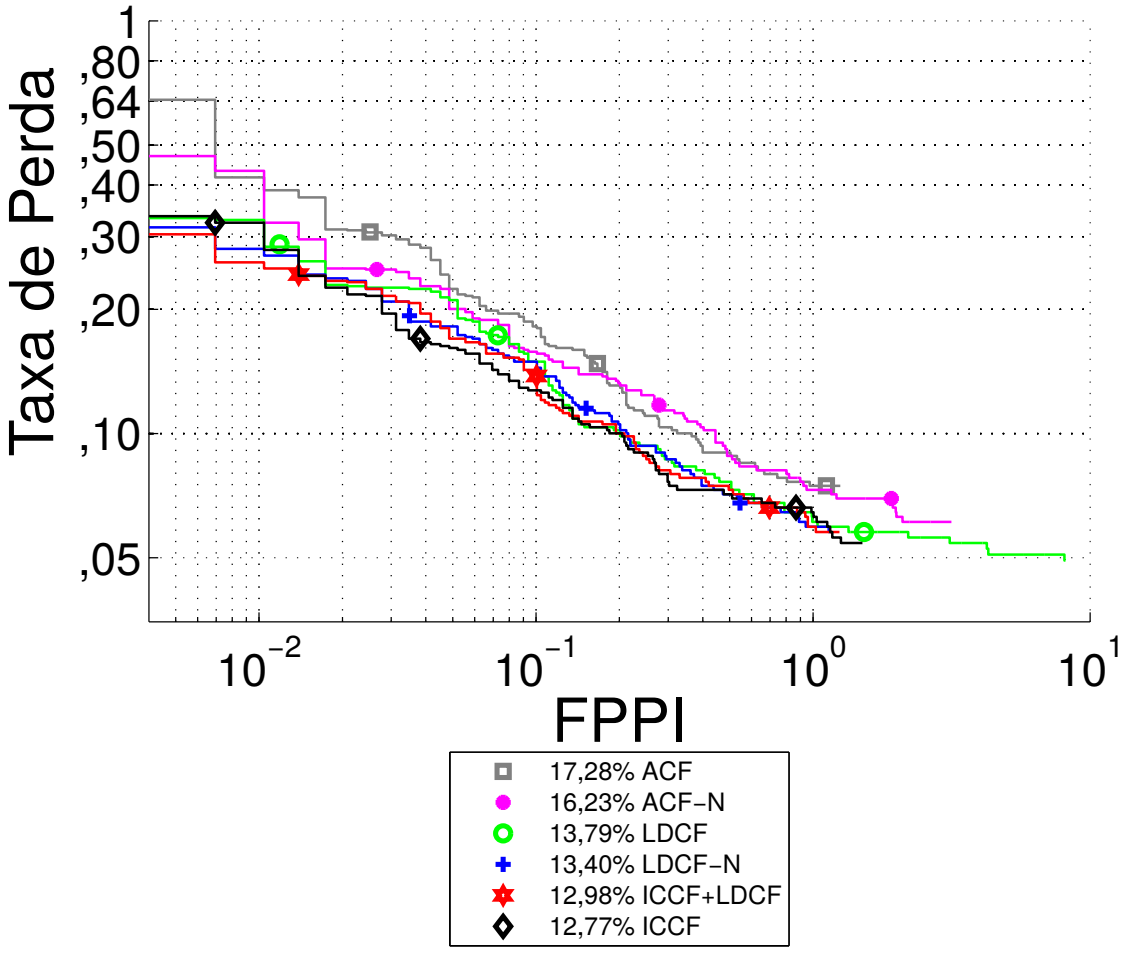
Por fim, por questões de completude, o Gráfico 3 apresenta os melhores resultados, dentre todas as rodadas de experimentos realizadas, dos detectores analisados para o Modelo 1. Os resultados das versões originais dos detectores LDCF e ACF são também apresentados para referência. Esses detectores também empregam a base de dados INRIA como base de dados de validação. ACF-N (semente aleatória 28) e LDCF-N (12 filtros LDCF - semente aleatória 55) foram treinados e testados na mesma máquina que os detectores ICCF+LDCF (18 filtros ICCF e quatro filtros LDCF - semente aleatória 92) e ICCF (25 filtros ICCF - semente aleatória 28). Os resultados originais dos detectores ACF e LDCF estão disponíveis publicamente em (DOLLÁR, 2016).

Base de dados Caltech

Modelo 2

Neste experimento, a avaliação é conduzida na base de dados Caltech usando, principalmente, os parâmetros do processo de extração de características levantados na base de dados de validação INRIA. A Tabela 5 apresenta os resultados para o referido conjunto de parâmetros. Repare que o esquema de amostragem Caltech-7,5Hz é utilizado no processo de treinamento do Modelo 2.

Gráfico 3 – Comparação entre diferentes detectores (INRIA). A LogAvrMR de cada método é apresentada na legenda.



Devido aos maiores tempos demandados pelas etapas de treinamento e detecção, apenas cinco rodadas com diferentes sementes aleatórias foram realizadas. Esse maior tempo de treinamento se deve ao maior número de árvores empregadas, que também possuem maior profundidade máxima, e a maior quantidade de exemplos negativos e positivos utilizados no treinamento (Tabela 1).

Tabela 5 – Desempenho dos detectores ACF, LDCF, ICCF e ICCF+LDCF (Caltech-7,5Hz).

Configuração	LogAvrMR ± desvPad (%)
ACF	29,30 ± 0,63
LDCF-4	25,04 ± 0,57
LDCF-12	26,07 ± 0,69
ICCF-18	25,98 ± 0,65
ICCF-32	25,70 ± 0,75
ICCF+LDCF (22)	**24,82 ± 0,48**
ICCF+LDCF (30)	25,05 ± 0,65

Repare que o valor médio da LogAvrMR relativa ao LDCF é ligeiramente menor que o melhor resultado apresentado pelo ICCF. No entanto, em termos de desempenho médio, para um maior número de filtros, o ICCF demonstra um aumento de desempenho, enquanto o LDCF

apresenta uma redução. É importante notar que, para o Modelo 2, o ICCF empregando 32 filtros possui um desempenho médio superior ao que utiliza 18 filtros. Isso vai de encontro ao observado para o Modelo 1 na base de dados INRIA. Conforme hipótese considerada na análise realizada naquela base de dados, empregar mais filtros com um modelo de maior capacidade e que utiliza mais dados no treinamento parece realmente ser mais adequado. Quando se analisa também o desvio padrão dos experimentos, percebe-se que os detectores ICCF e LDCF possuem qualidade de detecção dentro do intervalo de variação da LogAvrMR. Isso indica uma competitividade entre essas duas abordagens, em termos da métrica utilizada, também na base de dados Caltech, para o Modelo 2. O resultado para o detector ACF também é exibido e ele apresenta desempenho inferior aos detectores ICCF e LDCF que, conforme mencionado, apresentam-se compatíveis em termos de qualidade de detecção.

Novamente, como um indicativo de complementariedade entre as duas metodologias, o melhor resultado é alcançado para o detector ICCF+LDCF, usando 22 filtros (18 filtros ICCF e quatro filtros LDCF - 22 no total). A partir da Tabela 5, pode-se também notar que, mesmo para um modelo de maior capacidade de aprendizado, aumentando o número de filtros LDCF para 12 não resulta em uma melhora no desempenho do detector ICCF+LDCF (18 filtros ICCF e 12 filtros LDCF - 30 no total). Isso pode ser causado pela redução no desempenho médio do LDCF-12, que era superior ao relativo a quatro filtros, na base de dados INRIA, mas que na base de dados Caltech é inferior.

Modelo 3

O detector ICCF também é avaliado junto ao mais robusto Modelo 3. Conforme pode ser visto na Tabela 1, esse modelo emprega mais exemplos negativos e possui um processo de busca exaustiva no conjunto de características nas primeiras 512 árvores do último estágio de treinamento. Os Modelos 2 e 3 possuem a mesma quantidade máxima de parâmetros a serem estimados. São 4096 árvores binárias com profundidade máxima igual a 5 (15 nós e 16 folhas). Logo, no máximo, $4096 \times 31 = 126.971$ parâmetros podem ser estimados. Contudo, o Modelo 3, por utilizar mais exemplos negativos, possui mais dados que o Modelo 2 para estimar esses parâmetros. Uma outra vantagem do Modelo 3, em relação ao Modelo 2, é o processo de aumento de dados de treinamento, conforme mencionado na Seção 2.3.1.

A Tabela 6 apresenta os desempenhos dos detectores ACF, LDCF e ICCF, para o Modelo 3. A média e o desvio padrão, relativos a cinco rodadas com diferentes sementes aleatórias, são apresentados. Repare que para essa configuração de parâmetros, que permite uma maior capacidade de aprendizado por parte do classificador, o valor médio da LogAvrMR relativa ao ICCF (18 filtros) é menor que a apresentada pelo LDCF (4 filtros). Esse fato não foi observado, por exemplo, para os Modelos 1 e 2. O maior número de características do ICCF, resultado do emprego de uma maior quantidade de filtros, aparenta ser melhor aproveitado pelo Modelo 3. No que tange ao processo de extração de características, os parâmetros utilizados para o ICCF (18 filtros de dimensões 5×5) são aqueles obtidos na base de dados INRIA. O detector LDCF

emprega o conjunto de parâmetros originais (4 filtros de dimensões 5×5), obtidos em (NAM; DOLLÁR; HAN, 2014).

Tabela 6 – Desempenhos médios do ACF, LDCF e ICCF (Caltech-7,5Hz).

Detector	ACF	LDCF	ICCF
LogAvrMR ± desvPad (%)	20,17±0,68	17,36±0,57	16,26±0,52

Assim como apresentado na literatura, o Gráfico 4 exibe uma comparação entre os melhores resultados de diferentes detectores na base de dados Caltech. Os resultados das detecções de cada um dos métodos está disponível em (DOLLÁR, 2016), mas nenhuma informação sobre as sementes aleatórias empregadas nos experimentos é fornecida. No Gráfico 4, o LDCF++ é o único detector não baseado em DCNN com um desempenho superior ao ICCF (semente aleatória 92), embora a diferença seja por uma margem relativamente pequena. Além disso, o LDCF++ é uma versão modificada do detector LDCF, que emprega um classificador SVM adicional às árvores de decisão. Esse classificador não é da família de canais de características filtradas (FCF) e é treinado com base no contexto ao redor das janelas de detecção, obtidas ao se aplicar o detector LDCF no conjunto de treinamento. O classificador SVM é então utilizado para modificar os graus de confiabilidade das detecções realizadas pelo LDCF, conforme descrito em (OHN-BAR; TRIVEDI, 2016). Essa é uma contribuição ortogonal à apresentada neste capítulo, podendo também ser agregada de forma incremental ao detector ICCF. Contudo, a inserção de um classificador adicional tende a aumentar o custo computacional do processo de detecção, fato esse não abordado em (OHN-BAR; TRIVEDI, 2016). O resultado apresentado para o LDCF++ foi extraído de (OHN-BAR; TRIVEDI, 2016) e os resultados apresentados para o ACF (semente aleatória 0) e LDCF (semente aleatória 82) foram obtidos por meio de experimentos realizados na mesma máquina que o detector ICCF.

O Gráfico 5 e a Figura 10 apresentam, respectivamente, o percentual de características de cada canal e a distribuição espacial das características selecionadas pela AdaBoost, para o detector ICCF. Percebe-se, por meio do Gráfico 5, uma maior importância no processo de decisão por parte das características derivadas a partir do canal L. No entanto, novamente, os canais possuem uma frequência de utilização semelhante. Por meio da Figura 10, é possível notar uma boa distribuição espacial das características selecionadas pela AdaBoost, de maneira que o modelo do pedestre apresenta uma silhueta próxima a de um ser humano. Isso indica a importância da informação espacial na tarefa de classificação.

Por fim, a Tabela 7 apresenta os melhores resultados dos detectores ACF, LDCF e ICCF, dentre cinco rodadas com diferentes sementes aleatórias para o Modelo 3, porém considerando agora o conjunto mais preciso de anotações da base de dados Caltech (Caltech-10Hz$_{PREC}$). Esse conjunto de dados apresenta um procedimento de marcação das janelas que inscrevem melhor o pedestre, evitando excesso de fundo da cena. Dessa forma, um melhor conjunto de janelas (*bounding boxes* - BB) é utilizado no treinamento. Observe que os três detectores apresentam um ganho considerável de rendimento. O ICCF, novamente, no que tange a LogAvrMR, mostrou-se

Gráfico 4 – Comparação entre diferentes detectores (Caltech-7,5Hz). A LogAvrMR de cada método é apresentada na legenda.

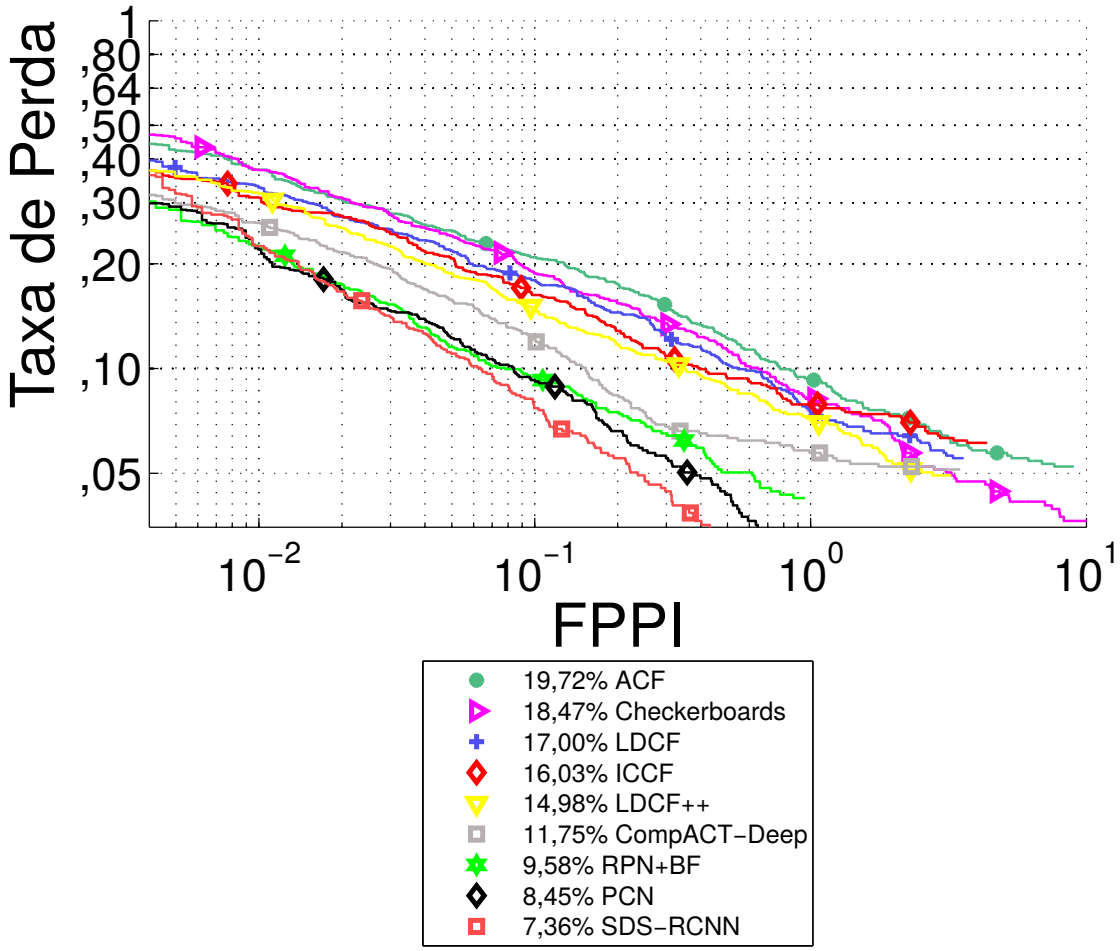
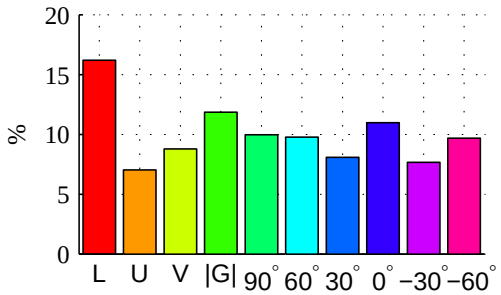
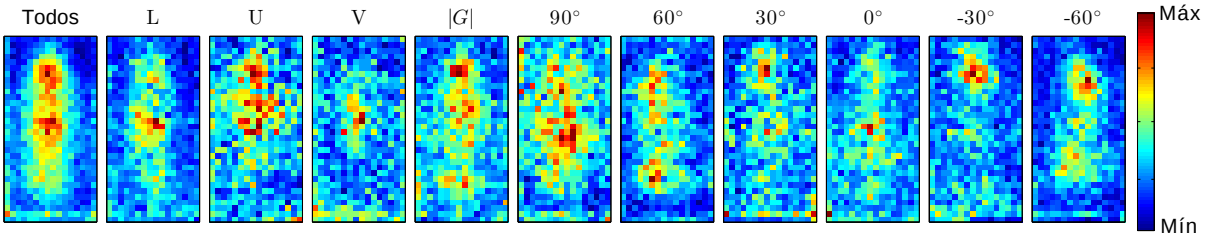


Gráfico 5 – Percentual de características selecionadas a partir de cada canal pela AdaBoost (Caltech-7,5Hz).



competitivo com os demais membros da sua família de soluções. A Tabela 7 apresenta ainda os melhores resultados, no conjunto de dados Caltech-10Hz$_{PREC}$, dos detectores LDCF++ e RotatedFilters+VGG16 (ZHANG et al., 2016b), extraídos de (OHN-BAR; TRIVEDI, 2016). Esse último possui, em seu fluxo de operação, a DCNN intitulada VGG16. Observe que o detector ICCF apresenta desempenho próximo ao do detector RotatedFilters+VGG16, mesmo ele se valendo de uma solução baseada em aprendizado profundo. Repare ainda que os detectores ICCF e LDCF++, para o conjunto de dados Caltech-10Hz$_{PREC}$, apresentam uma menor diferença,

Figura 10 – Distribuição das posições das características selecionadas pela AdaBoost nas cinco rodadas do detector (Caltech-7,5Hz). As cores presentes na figura representam o grau de utilização, de cada posição do modelo do pedestre, no processo de decisão. Nesse sentido, as siglas Máx e Mín representam os valores relativos às utilizações máxima e mínima, respectivamente.



entre um e outro, em relação ao conjunto Caltech-7,5Hz (Gráfico 4). Como os pedestres do conjunto Caltech-10Hz$_{PREC}$ são melhores inscritos nas BB, a análise de contexto realizada pelo classificador SVM adicional, empregado pelo LDCF++, parece ser menos determinante nesse caso.

Em uma análise um pouco mais desbalanceada, os detectores LDCF, LDCF++ e ICCF também apresentam LogAvrMR próximas a da solução baseada em aprendizado profundo RPN+BF, apresentada no Gráfico 4, quando essa não tem acesso ao conjunto mais preciso de anotações. Contudo, é fato que essa análise é realizada em um cenário favorável ao LDCF e ICCF, pois é esperado que o RPN+BF também apresente um incremento na qualidade da detecção ao empregar esse melhor conjunto de dados anotados. Essa análise visa apenas demonstrar que, mesmo não utilizando aprendizado profundo, os conjuntos de características do LDCF e ICCF podem ser competitivos com soluções baseadas em DCNN. Para isso, outros recursos podem ser utilizados, como o emprego de um melhor conjunto de dados anotados. É importante frisar também que as soluções RPN+BF (ZHANG et al., 2016a), SDS-RCNN (BRAZIL; YIN; LIU, 2017) e PCN (WANG et al., 2017) possuem a vantagem, em relação aos métodos analisados neste trabalho, de empregarem modelos pré-treinados na base de dados ImageNet (DENG et al., 2009) (milhões de imagens). Esse fato também constitui um desbalanceamento a favor dos métodos baseados em aprendizagem profunda, em relação aos detectores LDCF e ICCF. Por fim, os detectores SDS-RCNN e PCN empregam ainda segmentação semântica e análise de contexto, respectivamente, como informações adicionais no processo de detecção.

Tabela 7 – Desempenho dos detectores de pedestres para o conjunto Caltech-10Hz$_{PREC}$.

Detector	ACF	LDCF	ICCF	LDCF++	RotatedFilters+VGG16
LogAvrMR	12,40	10,31	9,89	9,71	9,32

Análise de tempo de detecção

A Tabela 8 apresenta as taxas de detecção, em FPS, dos métodos ACF, LDCF e ICCF. Em (DOLLÁR et al., 2014; CAO; PANG; LI, 2016), os tempos de detecção dos detectores ACF e

LDCF foram avaliados na base de dados INRIA. No entanto, a Tabela 8 apresenta essa informação para os três modelos empregados ao longo dos experimentos. É importante ressaltar que, para uma mesma quantidade de filtros e o mesmo classificador, o custo computacional do LDCF e ICCF são iguais durante o processo de detecção. Isso ocorre devido ao fato do fluxograma de operação ser o mesmo, baseado no processo de filtragem dos canais HOG+LUV. No entanto, conforme pode ser visto na referida tabela, por empregar 18 filtros, o ICCF possui uma menor taxa de FPS que a versão original do LDCF, que emprega 4 filtros. Sendo assim, embora o ICCF supere o LDCF em termos de LogAvrMR, em alguns cenários, o LDCF possui como ponto positivo o tempo de detecção, que é inferior. Na análise de tempo realizada, foi utilizado um computador com processador Intel® Core^{TM} i7-960, 3.20 GHz e 24GB de memória RAM.

Tabela 8 – Taxas de detecção, em FPS, dos detectores ACF, LDCF e ICCF.

Modelo	ACF (FPS)	LDCF (FPS)	ICCF (FPS)
1	13,50	3,25	0,85
2	10,20	2,85	0,80
3	2,70	1,60	0,65

2.4 Conclusões deste capítulo

Neste capítulo, realizou-se uma avaliação do uso da análise de componentes independentes (ICA) no procedimento de extração de características a partir dos canais HOG+LUV. Foram utilizados no processo de comparação dois detetores da família de características de canais filtrados (FCF), denominados ACF e LDCF. Os detectores dessa família de soluções são elegíveis para serem empregadas no serviço de detecção de seres humanos proposto no Capítulo 3. Isso se deve ao fato desses detectores apresentarem, em geral, baixo custo computacional, não demandando, portanto, o uso de GPU. Além disso, esses métodos apresentam bons desempenhos, em termos de qualidade de detecção, conforme apresentado em (OHN-BAR; TRIVEDI, 2016). O detector proposto nesta tese, denominado ICCF, também é membro da família de soluções de FCF.

De forma a limitar a análise apenas ao procedimento de extração de características proposto, foram utilizados três modelos de classificadores com parâmetros avaliados na literatura. Esses modelos possuem diferentes capacidades de aprendizado, de forma que cada um utiliza de maneira diferente as características disponibilizadas. Ao longo dos experimentos realizados, avaliando-se os desempenhos médios e também os melhores resultados, percebeu-se que o ICCF possui melhor qualidade de detecção que o LDCF em alguns cenários. No entanto, ao se analisar o desvio padrão das rodadas de experimentos, conclui-se que o ICCF é um detector competitivo com o LDCF e que possui um melhor desempenho que o ACF em termos da métrica LogAvrMR. No âmbito da base de dados Caltech, ao se empregar o Modelo 3, sugerido em (OHN-BAR; TRIVEDI, 2016), em conjunto com anotações mais precisas dessa base de dados, o detector ICCF, assim como o LDCF, apresenta uma redução significativa na LogAvrMR, aproximando-se dos resultados de detectores baseados em DCNN.

3 Detecção de seres humanos no contexto de espaços Inteligentes

Conforme apresentado no Capítulo 2, a detecção de seres humanos é uma área de estudo muito ativa na comunidade científica. Mesmo com o sucesso da detecção geral de objetos (GIRSHICK et al., 2014; CAI et al., 2016; REN et al., 2017; REDMON; FARHADI, 2016), a detecção de seres humanos tem sido tratada de forma independente. Isso se deve, principalmente, ao fato de seres humanos serem componentes-chave em diferentes aplicações, conforme mencionado em (ZHANG et al., 2016a). Além disso, a detecção de seres humanos apresenta problemas específicos, no que diz respeito a discriminabilidade em relação ao fundo da imagem, principalmente em baixa resolução, conforme indicado em (MAO et al., 2017).

Nos últimos anos, redes neurais convolucionais profundas (DCNN) contribuíram, significativamente, para a melhoria da qualidade da detecção de objetos e humanos em imagens. No entanto, mesmo os melhores detectores genéricos apresentam, ainda, limitações ao se considerar a generalização para diferentes bases de dados, conforme constatado em (ZHANG; BENENSON; SCHIELE, 2017). Isso indica, assim como o estudo realizado em (LI; YAO; WANG, 2012), que o uso de informação adicional pode ser necessário. O projeto de soluções específicas, que consideram ambientes parcialmente estruturados, pode ainda ser inevitável em algumas situações.

Neste capítulo, agregam-se diferentes conceitos como homografia e segmentação de imagens a dois detectores de pedestres, de forma a se compor um serviço de detecção de seres humanos. Esse serviço é oferecido em um espaço inteligente, que usa uma rede de câmeras como principal forma de sensoriamento. Diferentemente de soluções que empregam DCNN, os detectores utilizados no serviço desenvolvido não dependem de *hardware* específico, como unidades de processamento gráfico (*graphical processing units* - GPU). Esse fato torna as tarefas de alocação e distribuição de serviços, em nós de processamento da infraestrutura, mais simples. Isso é verdade, principalmente, para situações nas quais nem todos os nós da infraestrutura possuem GPU. Além disso, os detectores empregados apresentam uma boa relação custo computacional versus qualidade de detecção. Essas características tornam o serviço desenvolvido ao mesmo tempo eficiente e flexível.

Em (RIBEIRO et al., 2017), por exemplo, diferentemente do sistema desenvolvido nesta tese de doutorado, o uso de GPU é mandatório para que a detecção de seres humanos seja executada no contexto da aplicação desenvolvida. Além disso, o procedimento de detecção implementado no referido trabalho não é entregue para o espaço inteligente como um serviço. Logo, ele não pode ser flexivelmente utilizado por diferentes aplicações e instanciado em diferentes nós da infraestrutura, em qualquer tempo. Conceitos relacionados a sincronismo, escalabilidade e confiabilidade do serviço de detecção não são abordados. Emprega-se um nó de processamento

com alta capacidade computacional (GPU), de maneira a atingir uma baixa taxa de falsos positivos. Nesta tese, por outro lado, empregam-se técnicas simples e a redundância provida por uma rede de câmeras para atingir esse mesmo objetivo.

Por fim, é importante mencionar que o serviço desenvolvido neste trabalho não está impossibilitado de utilizar GPU. No entanto, o objetivo é desenvolver uma solução que não gere restrições na infraestrutura, que seja adequada para diferentes aplicações que empreguem uma rede de câmeras e que evite o uso de soluções computacionalmente custosas.

3.1 Trabalhos relacionados

3.1.1 A detecção de seres humanos em espaços inteligentes

Nesta seção, serão discutidos alguns trabalhos da literatura de espaços inteligentes, que abordam a detecção de seres humanos utilizando redes de sensores. Alguns diferentes tipos de sensores podem ser abordados, mas o foco são trabalhos que empregam sensores visuais. A ideia principal é apresentar as diferenças entre a solução desenvolvida nesta tese e os trabalhos que abordam tema semelhante na literatura.

Em (SURIE; PARTONIA; LINDGREN, 2013), um sensor Kinect$^{\mathrm{TM}}$, montado em uma parede, foi utilizado para detectar seres humanos. Para este fim, foram empregadas a imagem RGB e o campo de profundidade do sensor Kinect$^{\mathrm{TM}}$, de forma a realizar detecção de esqueleto e reconhecimento de faces. Entretanto, no presente trabalho, o serviço de detecção de seres humanos emprega um modelo baseado em aparências e depende apenas de imagens RGB. Além disso, nenhum sensor telêmetro é empregado para obter informação 3D.

Os autores em (SURIE; PARTONIA; LINDGREN, 2013) também mencionam que o espaço de trabalho da solução proposta é restrito, devido ao único e limitado campo de visão da câmera do sensor Kinect$^{\mathrm{TM}}$. Eles mencionam inclusive que a solução desenvolvida é estruturada, podendo apresentar falhas para um campo de visão expandido. Por outro lado, a arquitetura da solução, proposta nesta tese, se beneficia dos múltiplos pontos de vista disponibilizados por uma rede de câmeras. Além disso, o detector, que é proposto como um serviço, é capaz de identificar seres humanos em diferentes poses em relação a esses sensores. A única configuração demandada *a priori* é a calibração do sistema de câmeras.

Em (GLAS et al., 2013), desenvolve-se um "sistema de robôs em rede" ("*network robot system*") de forma a implantar robôs sociais em aplicações práticas em *shoppings* e outros espaços comerciais. No referido trabalho, aplicações de ambientes inteligentes e interação homem-máquina são implementadas usando dados obtidos de sensores instalados no espaço de trabalho. Além do uso de câmeras, sensores laser, telêmetros e leitores de impressões digitais são utilizados e instalados no ambiente. Um supervisor humano também é considerado quando o sistema encontra dificuldades em tarefas de reconhecimento de voz e de pessoas, além do planejamento de caminho.

No sistema apresentado nesta tese de doutorado, nenhum supervisor humano é empregado e apenas câmeras são utilizadas. O emprego apenas de sensores visuais é preferível, haja vista que esses sensores já se encontram disponíveis na maior parte dos ambientes comerciais e públicos. Além disso, em (GLAS et al., 2013), o sistema não é descrito como sendo implementado em uma arquitetura de *software* baseada em serviços. O conceito de serviço está limitado ao de tarefas que podem ser executadas pelos robôs para os seres humanos. Já na arquitetura da plataforma aqui apresentada, o serviço de detecção de seres humanos proposto é um dos diferentes serviços (atuadores dos robôs, câmeras) disponibilizados para as aplicações. A interação entre esses serviços é projetada para ser harmoniosa, de maneira a garantir o sucesso das aplicações atendidas.

Um sistema robótico de transporte para assistência a compras é desenvolvido em (MATSUHIRA et al., 2010). Diferentemente do trabalho desenvolvido nesta tese, o processo de detecção de seres humanos não é apresentado como um serviço que utiliza conceitos de computação em nuvem. Por isso, a solução proposta não é apresentada como sendo flexível, de maneira a poder ser instanciada em qualquer nó da infraestrutura. O sistema de detecção aparenta ser implantado de forma rígida, como uma aplicação que é executada em nós atrelados, fisicamente, a cada uma das câmeras empregadas na infraestrutura. A aplicação desenvolvida, apesar de interessante, possui um sistema simples de detecção de seres humanos, baseado apenas em técnicas de processamento de imagens e subtração de fundo. Além disso, o modo de seguimento de pessoas depende também de um sensor laser de forma a ser funcional. Apesar do emprego de um sistema de câmeras, a informação 3D não é recuperada como no sistema apresentado nesta tese, em que informações de calibração e homografia são utilizadas.

O objetivo em (ALBAWENDI et al., 2015) é investigar um sistema de monitoramento visual de baixo custo, de forma a assistir idosos que demandem cuidados especiais. A ideia é limitar a quantidade de informação transmitida pelo sensor visual, reduzindo o nível de intrusão da câmera na rotina dos usuários. O sistema desenvolvido no trabalho funciona mais como uma aplicação e o procedimento de detecção de objetos é muito dependente de subtração de fundo, requerendo um espaço de trabalho mais estruturado que o apresentado na solução desta tese de doutorado.

Em (ADDUCI; AMPLIANITIS; REULKE, 2014), uma aplicação de rastreamento e detecção 3D de seres humanos, empregando uma rede de câmeras, é construída. No entanto, de acordo com os autores, a aplicação desenvolvida necessita de modificações de *hardware* e *software* para ser preparada para tarefas de tempo real. O serviço de detecção de seres humanos proposto nesta tese, bem como os de detecção e de controle de robôs empregados, é adequado para aplicações de tempo real. Além disso, diferentemente dos trabalhos citados anteriormente, os serviços são ofertados utilizando conceitos de computação em nuvem. O objetivo é que os serviços atendam, nesse paradigma, aos requisitos de tempo e confiabilidade das aplicações. Tudo isso é considerado no projeto do fluxograma do detector de seres humanos.

Os autores de (LEE et al., 2012) desenvolveram um sistema baseado em visão para realizar a

localização de seres humanos e robôs. Esse sistema é então utilizado em um serviço de display ativo em um espaço inteligente. Para esse fim, configura-se uma rede de câmeras para estimar as posições tridimensionais de seres humanos e robôs, que são detectados utilizando características obtidas a partir do HOG. Nesse trabalho, diferentemente da solução aqui desenvolvida, a arquitetura não apresenta serviços flexíveis, passíveis de serem instanciados em toda a infraestrutura, em qualquer tempo. Embora se utilizem diversos dispositivos inteligentes de rede distribuídos (*distributed intelligent network devices* - DIND), que consistem em uma câmera, um processador e um dispositivo de rede, o sistema possui uma configuração rígida, com nós com funções especializadas.

Além disso, em (LEE et al., 2012), de forma a se estimar as posições tridimensionais de seres humanos, se faz necessário realizar a correspondência entre pelo menos duas detecções providas por diferentes câmeras. Dessa forma, um procedimento de triangulação deve ser realizado. A função de localização 3D do referido trabalho pode apresentar problemas se muitas pessoas estão presentes no espaço de trabalho, tendo em vista que ter duas detecções de uma mesma pessoa é mandatório. Já no caso apresentado neste trabalho, dado que se assume que humanos e robôs estão sempre em contato com o plano do chão, o sistema pode obter informações tridimensionais apenas com uma imagem. Detecções adicionais são então utilizadas para remover falsos positivos e melhorar a estimativa da localização 3D. Finalmente, o espaço inteligente de (LEE et al., 2012) é mais estruturado, sem objetos na cena definida para os experimentos e com um fundo claro e homogêneo, que facilita a detecção de pessoas e robôs.

Em (AHMEDALI; CLARK, 2006), um sistema colaborativo de câmeras é construído para realizar a detecção de seres humanos. Contudo, o processo de detecção está embutido em microprocessadores atrelados às câmeras, o que torna a arquitetura menos flexível que a proposta nesta tese de doutorado. Em (WANG et al., 2012), uma arquitetura distribuída e escalável para aplicações de visão computacional é proposta. A ideia de empregar aplicações de visão computacional como serviços é similar em espírito à apresentada nesta tese. Contudo, apenas rotinas simples de visão computacional são implementadas, de forma que modelos obtidos por meio de aprendizado não são empregados. A avaliação da interação do serviço de rastreamento de *blobs* com outros serviços ou aplicações de tempo real não é apresentada. Além disso, não se menciona a existência, na arquitetura, de um controlador de replicação tal qual o Kubernets, que é empregado na proposta do presente trabalho. Por isso, aspectos relacionados a confiabilidade do sistema não são abordados. Diferentemente deste capítulo, em que um barramento especializado para troca de mensagens é empregado, em (WANG et al., 2012) um esquema menos flexível baseado em filas e memória compartilhada é utilizado.

Em (CANEDO-RODRIGUEZ et al., 2012), um sistema de câmeras inteligentes, com o objetivo de prover implantação rápida e fácil de robôs ubíquos, é proposto. O sistema é autoconfigurável e distribuído por meio de agentes de câmeras, que possuem capacidade própria de processamento. A flexibilidade do sistema parece estar relacionada a capacidade de autoconfiguração e, consequentemente, fácil mobilidade física dos agentes de câmeras. No entanto, não

se aborda aspectos relacionados a flexibilidade dos módulos implementados, internamente aos agentes de câmeras, na infraestrutura de software. Dessa forma, não fica claro se esses módulos são implementados como serviços, que podem ser dinamicamente alocados em qualquer ponto da infraestrutura existente, em qualquer tempo. Os agentes de câmeras são, inclusive, os responsáveis pela comunicação, possuindo um módulo específico para isso. Apenas a detecção do robô, por meio de um conjunto de diodos emissores de luz é realizada. O próprio processo de detecção do robô parece ser implementado como uma aplicação interna a cada agente de câmera. No trabalho apresentado nesta tese, as câmeras fazem parte da infraestrutura e são utilizadas como serviços, não possuindo recursos de processamento específicos atrelados a elas. As funções de detecção desenvolvidas são instanciadas em qualquer parte da infraestrutura, também como serviços, que se comunicam de forma independente.

Em (MATSUYAMA; UKITA, 2002), um sistema cooperativo distribuído de rastreamento é desenvolvido. Esse sistema é composto por agentes de visão ativos (AVA), que consistem de câmeras ativas com computadores conectados a rede. Cada AVA é constituído de módulos de percepção, ação e comunicação. Esses módulos são implementados como processos Unix e possuem um esquema complicado de comunicação, que emprega memória compartilhada. O sistema de percepção é baseado principalmente em subtração de fundo. Diferentemente, nesta tese, o serviço de detecção proposto emprega um barramento especializado para troca de mensagens e um processo de detecção mais sofisticado, que emprega um modelo baseado em aparências. Além disso, o serviço proposto neste trabalho é mais flexível e independente que os módulos internos aos AVA desenvolvidos em (MATSUYAMA; UKITA, 2002).

Em (MUJA et al., 2011), uma infraestrutura de detecção modular é desenvolvida. A ideia é intercambiar detectores de objeto ao longo de uma tarefa, empregando nós de um sistema de operação de robôs (*robot operating system* - ROS). Com isso, espera-se agregar robustez, velocidade e escalabilidade ao processo de detecção. Testes são conduzidos apenas em imagens estáticas, sem validação em tarefas de tempo real. Em (MEHMOOD, 2015), o problema de detecção de seres humanos em uma rede de câmeras é abordado. No entanto, o foco é a proposição de um detector genérico de seres humanos e testes apenas em bases de dados públicas são conduzidos. O detector proposto não é apresentado em uma arquitetura baseada em serviços, que atende a diferentes aplicações de tempo real. Por fim, existem na literatura diversos esforços para se construir arquiteturas distribuídas para aplicações relacionadas à internet das coisas (*internet of things* - IoT) e robôs de serviços, sendo alguns exemplos encontrados em (SARKAR et al., 2015; PYO et al., 2015).

3.1.2 Comentários adicionais

Na Seção 3.1.1, foram destacadas algumas diferenças entre a solução proposta nesta tese e alguns trabalhos relacionados presentes na literatura. Contudo, ressalta-se que, ao alcance da revisão bibliográfica realizada, não se tem conhecimento de um detector de seres humanos

desenvolvido como um serviço tal qual apresentado neste trabalho, considerando aspectos de implementação e usabilidade.

Alguns trabalhos apresentam detectores com alta qualidade de detecção em bases de dados públicas. Contudo, é importante mencionar que a solução proposta, especificamente neste capítulo, não tem por objetivo propor um detector estado da arte em bases de dados específicas. Apesar da qualidade de detecção ser um fator considerado, o interesse principal é apresentar um serviço de detecção flexível e eficiente, que proveja detecção de seres humanos de forma distribuída para diferentes aplicações implantadas em um espaço inteligente. Nesse sentido, a escolha por detectores simples de seres humanos, em termos de custo computacional, permite que o serviço proposto seja oferecido para aplicações de tempo real. Isso também permite a utilização de nós de processamento regulares e não somente estações de trabalho com alto poder de processamento que, em geral, são dotadas de GPU.

Adicionalmente, paralelismo e escalabilidade podem ser encontrados em alguns trabalhos, normalmente quando um processador de múltiplos núcleos ou uma GPU é empregado para implementar um detector em um único nó. No entanto, no presente trabalho, a preocupação é o desenvolvimento de um serviço que possa ser distribuído sobre diferentes nós da infraestrutura do espaço inteligente, de acordo com a disponibilidade e capacidade computacional dos nós. Dessa forma, de acordo com a demanda, o serviço de detecção de seres humanos pode ser instanciado em diversos nós ou mesmo reiniciado e realocado se um nó se tornar indisponível. Essa abordagem possibilita um uso mais eficiente dos nós da infraestrutura, ao mesmo tempo que agrega confiabilidade ao processo. O paralelismo é intrínseco à arquitetura do sistema e ao fluxograma da solução.

Por fim, além das contribuições já mencionadas, o trabalho aqui desenvolvido permite realizar um paralelo entre aplicações restritas a testes em bases de dados públicas e aquelas empregadas em situações reais.

3.2 Solução proposta

Nesta seção, o serviço de detecção de seres humanos proposto é apresentado. De forma a situar melhor em qual contexto essa proposta se insere, são discutidos, a seguir, alguns conceitos relacionados à arquitetura de um espaço inteligente.

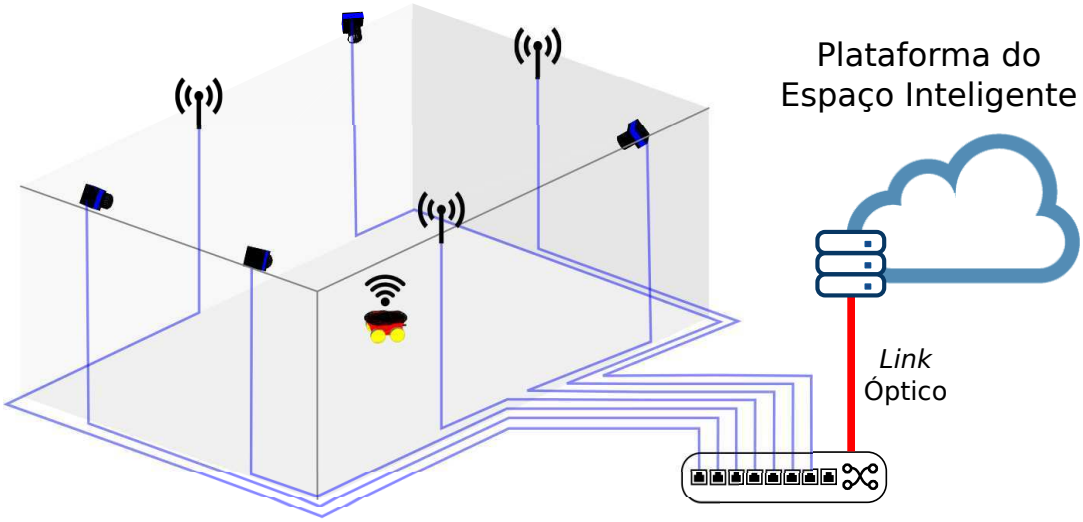
3.2.1 A arquitetura do espaço inteligente

Existem muitas definições para espaços inteligentes (WRIGHT; STEVENTON, 2004; LEE; HASHIMOTO, 2002). A definição adotada, nesta tese, é que um espaço inteligente pode ser descrito como um ambiente interativo, equipado com uma rede de sensores aptos a capturarem informações sobre o local analisado. Além disso, esse ambiente possui uma rede de atuadores, que podem ser diretamente controlados por serviços de forma a realizar modificações ou intervir

no ambiente. Além de controlar os atuadores, os serviços podem analisar a informação extraída por meio dos sensores, de maneira a suportar decisões e executar tarefas. Como sensores podem ser citadas câmeras, microfones e sensores de ultrassom e laser, enquanto entre os exemplos de atuadores estão robôs, dispositivos móveis e sistemas de saída de informação. Sensores, atuadores e serviços de computação são apoiados por uma infraestrutura de *software*, que é responsável por prover canais de comunicação e todas as demais abstrações necessárias. Serviços ou recursos de um dispositivo específico (sensores e atuadores) podem ser acessados e utilizados por diferentes entidades, tais como outros serviços, aplicações ou mesmo outros dispositivos.

Conforme mencionado anteriormente, o espaço inteligente empregado neste trabalho é baseado em visão computacional. Dessa forma, esse espaço está instrumentado com uma rede de câmeras (*internet protocol* - IP), aptas a capturar imagens e vídeos digitais, conforme apresentado na Figura 11. As câmeras, tal qual desenvolvido em (RAMPINELLI et al., 2014), são os principais sensores utilizados. O sistema também está preparado para controlar um robô. De forma a ter um alto nível de compreensão sobre o ambiente, uma infraestrutura de *software* está disponível para capturar e analisar, em tempo real, as informações adquiridas pela rede de câmeras.

Figura 11 – Conceito de espaço inteligente.



A infraestrutura do espaço inteligente utilizado é concebida como uma plataforma de desenvolvimento, isto é, na forma de uma plataforma como serviço (*platform as a service* - PaaS). Dessa forma, desenvolvedores de aplicações podem fazer uso de diferentes serviços, mesmo que alguns deles, inicialmente, tenham sido desenvolvidos para uma aplicação em específico. Sendo assim, esses serviços devem ser flexíveis o suficiente para atender, ao mesmo tempo, requisitos específicos das aplicações, mas também prover um alto nível de abstração em termos de programação para os desenvolvedores. O modelo de arquitetura orientada a serviços (SOA) é utilizado na infraestrutura de *software* do espaço inteligente, de forma a prover a programabilidade e reusabilidade necessária a nível de serviço. Essas características fazem a construção e a implantação de aplicações mais fáceis para desenvolvedores. Além disso, a integração de novos serviços à plataforma, a cada aplicação desenvolvida, torna-se mais simples. A arquitetura do espaço

inteligente empregado foi inicialmente proposta em (QUEIROZ, 2016). O objetivo principal, no referido trabalho, era atender aplicações de robótica baseadas em visão computacional, por meio do emprego do paradigma de internet das coisas. Nesta tese, o espaço inteligente é expandido por meio da proposição de um serviço de detecção de seres humanos, que opera em tempo real e atende a diferentes aplicações.

A plataforma do espaço inteligente é implantada no topo de uma infraestrutura que emprega conceitos de computação em nuvem, tal qual uma infraestrutura como um serviço (*infrastructure as a service* - IaaS). O objetivo é atender requisitos específicos de aplicações de visão computacional, tais como baixa latência, ampla largura de banda e alta capacidade de processamento. A programabilidade dessa infraestrutura habilita a plataforma a atender aos rigorosos requisitos de aplicações de tempo real (GOMES et al., 2017). Essas características foram agregadas ao espaço inteligente no trabalho desenvolvido em (PICORETI, 2017). No referido trabalho, a virtualização, a nível de sistema operacional, permitiu a distribuição dos serviços sobre a infraestrutura de *hardware* do espaço inteligente de forma escalável e confiável. A referida virtualização é alcançada por meio da solução denominada Docker (MERKEL, 2014), conforme será apresentado adiante. O serviço de detecção de seres humanos, proposto nesta tese, também é desenvolvido empregando essa solução. A forma, como o serviço de detecção é implementado, possibilita a transferência das características mencionadas para as aplicações.

A Figura 12 exibe a arquitetura empregada pelo espaço inteligente, de forma a suportar aplicações de visão computacional. Esta plataforma possui quatro camadas: sensoriamento, comunicação, *middleware* e aplicação, que serão descritas nas próximas seções.

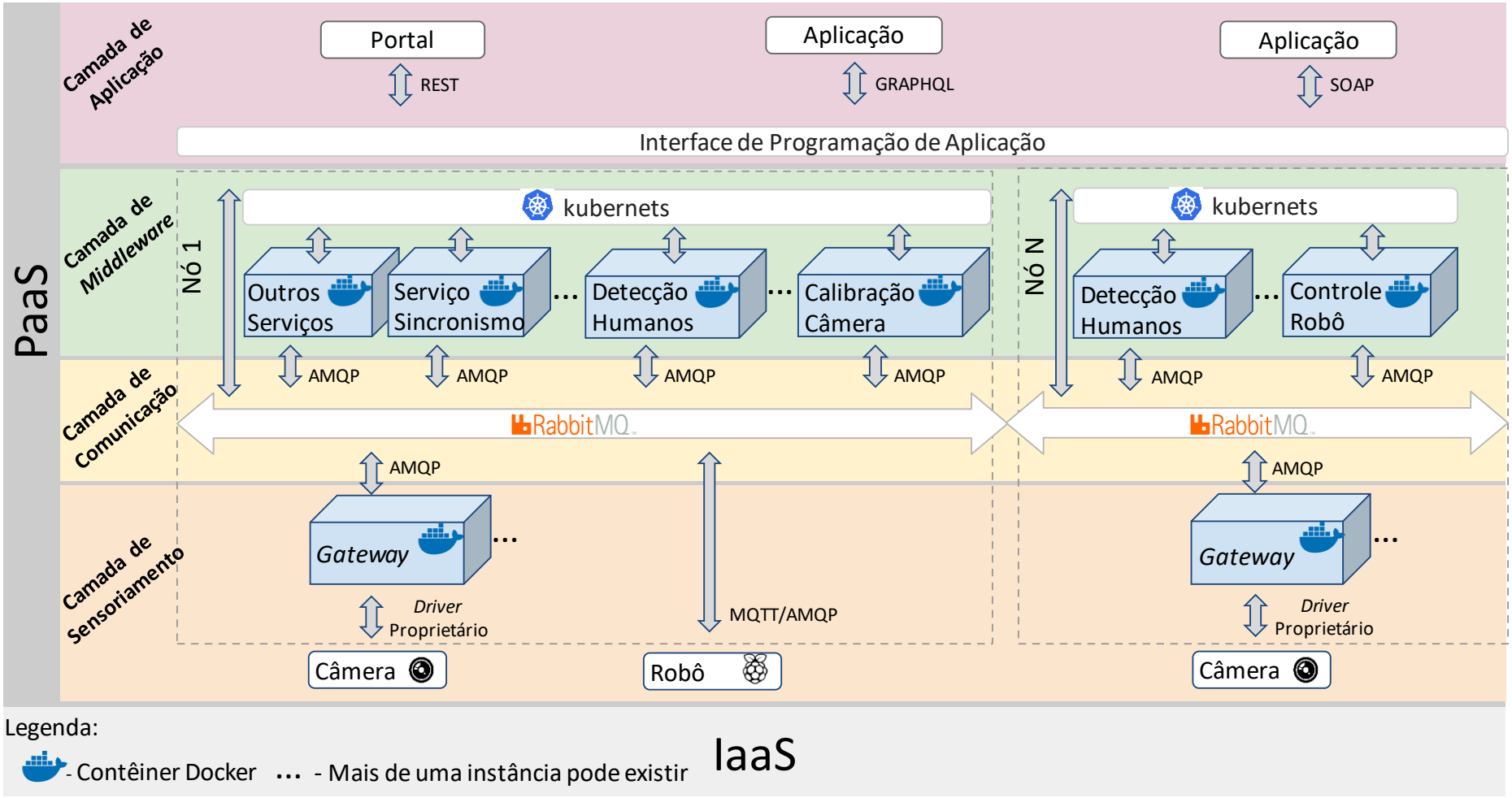
3.2.1.1 Camada de sensoriamento

A camada de sensoriamento é responsável pela exposição dos recursos do domínio físico para o digital. Além disso, essa camada deve simplificar a comunicação entre dispositivos heterogêneos por meio de interfaces padronizadas.

Para isso, essa camada adquire informação e controla o comportamento dos dispositivos no domínio físico. Como pode ser observado na Figura 12, as entidades físicas desta camada são sensores e atuadores. Cada entidade física no mundo real é representada por uma entidade virtual no domínio digital. A entidade virtual está associada a recursos que permitem uma interação, por meio de serviços, com a entidade física por ela representada.

Embora seja possível integrar diretamente equipamentos na plataforma, a função de padronização para esses equipamentos é normalmente realizada por meio de *gateways*. São esses elementos os responsáveis por traduzir o protocolo específico de um dispositivo para uma interface padrão disponibilizada pela entidade virtual. Da mesma forma, o *gateway* também é responsável pela conversão dos dados para um formato padrão.

Figura 12 – Arquitetura do espaço inteligente.



REST, GRAPHQL e SOAP - Ferramentas computacionais que podem ser empregadas na construção de aplicações e serviços *Web*;
AMQP (*advanced message queuing protocol*) - É um protocolo avançado de enfileiramento de mensagens;
MQTT (*message queue telemetry transport*) - É um protocolo que emprega o paradigma de publicação-inscrição (*publish-subscribe*) para a troca de mensagens.

3.2.1.2 Camada de comunicação

A camada de comunicação é responsável por rotear e encaminhar mensagens, realizar o desacoplamento de tempo e monitorar serviços. Para executar essas funções, toda a comunicação entre os componentes do sistema passa através de um intermediador (*broker*) de troca de mensagens. Na infraestrutura de *software* do espaço inteligente, empregado neste trabalho, a solução RabbitMQ é empregada como intermediador, conforme pode ser visualizado na Figura 12. Tendo em vista que todas as mensagens passam através do *broker*, é possível monitorar o fluxo de mensagens e até mesmo modificá-las antes de serem encaminhadas.

A principal vantagem de se ter uma plataforma que emprega um *broker*, para realizar a comunicação entre os elementos, é que uma entidade somente necessita saber o seu próprio endereço para se comunicar com as demais. Isso faz com que o desenvolvimento de serviços e aplicações seja mais simples. Além disso, existe a possibilidade de se ter mensagens persistentes. Se um destinatário encontra-se momentaneamente indisponível, a mensagem é armazenada até que ele esteja apto a recebê-la. O problema atual dessa abordagem é que o *broker* se torna um único ponto de falha e um gargalo em termos de desempenho. De forma a minimizar esses problemas, existe ainda a possibilidade de se empregar *brokers* clusterizados, que trabalham de forma integrada, mas em diferentes pontos da rede (ROSTANSKI; GROCHLA; SEMAN, 2014).

3.2.1.3 Camada de *Middleware*

O *middleware* provê interfaces de serviços para as aplicações, de forma a evitar que usuários tenham que lidar com aspectos não triviais das outras camadas. Essa camada é composta por diversos serviços que executam funções de suporte à infraestrutura e realizam rotinas específicas de visão computacional (Figura 12).

Dois importantes serviços de suporte à infraestrutura presentes na plataforma do espaço inteligente são o de sincronização e calibração das câmeras:

- Serviço de sincronização: normalmente, problemas de sincronismo em aplicações de visão computacional são comuns ao se trabalhar com duas ou mais câmeras em tempo real. A ausência de sincronismo pode gerar inconsistências entre as imagens capturadas, haja vista que não se pode garantir que as imagens correspondem a mesma cena e que foram tomadas no mesmo instante de tempo.

  O serviço de sincronização periodicamente monitora e aplica atrasos ao processo de captura das imagens das câmeras e, se necessário, a dados de outros dispositivos conectados, de forma a manter um erro máximo aceitável de sincronização dos dados para as aplicações desenvolvidas.

- Serviço de calibração das câmeras: este serviço implementa um procedimento semi-automático, conforme apresentado em (ZHANG, 2000), que retorna os parâmetros intrínsecos e extrínsecos das câmeras. Para isso, são necessárias a captura e o processamento de

imagens de um padrão quadriculado, que é movimentado manualmente por um ser humano. Uma vez que a calibração é executada, não existe mais a necessidade de recalibrar as câmeras, a menos que pelo menos uma delas seja deslocada da posição na qual a calibração tenha sido realizada.

No que diz respeito aos serviços específicos relacionados às rotinas de visão computacional, o foco são aqueles empregados nas tarefas apresentadas neste capítulo. O principal serviço é o de detecção de seres humanos, que é proposto nesta tese e apresentado na Seção 3.2.2. Além disso, serviços relacionados à detecção e controle do robô empregado, planejamento de caminhos e demais processos pertinentes são discutidos na Seção 3.2.3. Nessa mesma seção, é também abordado o inter-relacionamento entre os serviços e as aplicações.

Cada serviço na plataforma é virtualizado em um contêiner. Esse procedimento habilita o desenvolvimento de serviços fracamente acoplados, que são desenvolvidos de maneira independente. Muitos desses serviços podem ser conectados em um encadeamento de funções de serviços (*service function chaining* - SFC), de forma a compor uma aplicação. A virtualização de contêineres objetiva isolar os serviços e garantir independência, desde que as interfaces sejam mantidas. Esses contêineres podem ser facilmente compartilhados, implantados e atualizados. Além disso, eles podem ser escalonados instantaneamente e independentemente dos outros serviços que constituem a aplicação.

A solução conhecida como Docker é empregada como tecnologia de contêiner no serviço proposto nesta tese e também no espaço inteligente, tendo em vista que ela simplifica a construção de serviços internos aos contêineres. Além disso, a virtualização por meio de Docker é de baixo custo computacional e, por causa disso, múltiplas aplicações podem utilizar serviços ao mesmo tempo, no mesmo servidor físico ou virtual. Esse cenário habilita a um orquestrador prover uma quantidade adequada de recursos para os contêineres no momento certo. Dessa forma, permite-se uma melhor alocação dos contêineres em uma infraestrutura em nuvem, resultando em um melhor aproveitamento dos recursos. A orquestração dos contêineres Docker é realizada por meio da solução Kubernetes (BURNS et al., 2016), cujas funções incluem a implantação automática, escalonamento e operação de aplicações em contêineres.

3.2.1.4 Camada de aplicação

A última camada apresentada na Figura 12 é a de aplicação. Essa camada é a responsável por expor serviços, em alto nível e na forma de uma interface de programação de aplicação (*application programming interface* - API), para que desenvolvedores possam interagir com a plataforma. Essa API permite o desenvolvimento em diferentes linguagens de programação, deixando transparente o acesso aos serviços e equipamentos da plataforma.

Nessa camada, apenas alguns serviços são disponibilizados para os desenvolvedores. Normalmente, apenas aqueles necessários para a implantação de aplicações, tais como serviços de detecção e rastreamento de objetos e pessoas. Outros serviços, como os que suportam a

plataforma e executam a orquestração de recursos são ocultados, haja vista que não cabe ao usuário final se preocupar com questões relativas ao funcionamento da infraestrutura. O principal objetivo é ter uma plataforma o mais transparente quanto for possível para os usuários.

3.2.1.5 Características da Arquitetura

A plataforma, na qual o serviço de detecção de seres humanos é disponibilizado, apresenta uma série de características relativas ao domínio da aplicação. Todas essas características são transferidas pelo serviço proposto para as aplicações, sem que o desenvolvedor tenha que se preocupar com aspectos relacionados ao *hardware* ou *software* da plataforma. Sendo assim, o desenvolvedor pode concentrar todos os esforços na resolução do problema posto, utilizando serviços na construção da solução (aplicação). As seguintes características presentes na arquitetura do espaço inteligente podem ser destacadas:

- Escalabilidade: A plataforma precisa ser escalonável, de forma a atender a um aumento na demanda de recursos por parte das aplicações. Para um melhor gerenciamento dos recursos, os serviços devem ser desenvolvidos para serem simples, ter baixo acoplamento e serem implementados de uma forma que possam ser reutilizados por outras aplicações. Além disso, serviços desprovidos de estado e paralelizáveis podem ter novas instâncias levantadas pela plataforma para diferentes aplicações e em diferentes nós de processamento.

- Tempo real: A plataforma deve prover serviços de tempo real quando apenas a correta operação lógica de uma tarefa não é suficiente, mas também o seu tempo de execução. Como aplicações de visão computacional podem lidar com tarefas de tempo real, tendo algumas inclusive já sido listadas nesta tese de doutorado, entregar informações ou executar serviços para as aplicações no tempo demandado é algo crítico. Informações ou serviços atrasados podem tornar o sistema inutilizável ou até mesmo perigoso. Para isso, serviços de sincronismo e monitoramento rigorosos são primordiais.

- Confiabilidade: Aplicações devem se manter operacionais enquanto uma tarefa está sendo executada, mesmo na ocorrência de falhas na infraestrutura de *hardware* e *software*. Cada serviço que compõe uma aplicação, e que também suporta outros serviços, precisa ser confiável individualmente, de forma a garantir uma confiabilidade global do sistema. O espaço inteligente empregado tem mecanismos que permitem a distribuição, através da infraestrutura, de serviços de gerenciamento e suporte, além dos serviços específicos relacionados às aplicações. Em caso de falha de um nó de processamento, por exemplo, o serviço pode ser redistribuído, em tempo hábil, para outro ponto da infraestrutura.

De forma a atingir essas características, a arquitetura do sistema empregado é desenvolvida para permitir que os serviços sejam escalonados tanto verticalmente, quanto horizontalmente. O escalonamento vertical permite aumentar o poder computacional para uma mesma instância

de um serviço, enquanto o horizontal habilita o aumento do número de instâncias do mesmo serviço, de forma a lidar com um aumento da demanda.

Um exemplo, de como esses conceitos estão agregados ao serviço de detecção de seres humanos, é analisar a situação em que o número de câmeras no espaço inteligente aumenta. Conforme será explicado em detalhes na Seção 3.2.2, o serviço de detecção proposto está dividido em serviços menores. De forma resumida, a primeira etapa desse serviço consiste em um detector de seres humanos de baixo custo computacional, que atua no plano da imagem de cada uma das câmeras. Dessa forma, para cada câmera no ambiente, uma instância do serviço de detecção é inicializada. Haverá tantas instâncias quanto for a quantidade de câmeras e todas são executadas em paralelo.

Cada instância do serviço de detecção, da primeira etapa, provê um conjunto de detecções para um serviço subsequente, que filtra e agrupa as informações advindas de cada uma das instâncias dessa primeira etapa. Sendo assim, se o número de câmeras aumenta, o sistema irá executar um escalonamento horizontal do serviço de detecção, de forma a aumentar o número de instâncias ao máximo possível que cada nó de processamento pode lidar. Uma vez que o limite máximo é atingido, o sistema inicia novas instâncias do serviço em diferentes nós computacionais. Dessa forma, o grande volume de tráfego de dados, resultado do aumento do número de câmeras, será distribuído ao longo da infraestrutura, não gerando um gargalo de tráfego em um único ponto. Isso torna a operação do sistema viável.

No que diz respeito a confiabilidade, ela é garantida na arquitetura do sistema por meio do controlador de replicação denominado Kubernetes (BURNS et al., 2016). Esse controlador mantém o número desejado de instâncias de um serviço rodando. Se um nó da infraestrutura torna-se indisponível, os serviços que estavam sendo executados nesse nó de processamento são reinicializados em outros nós da infraestrutura. Dessa forma, o fato do detector de seres humanos ser implementado como um serviço, faz com que ele herde todas essas características importantes providas pelo paradigma no qual o espaço inteligente está inserido. A confiabilidade é especialmente importante, por exemplo, para os serviços de filtragem de janelas (*bounding box* - BB) de detecção e controle de robôs, que serão explicados nas Seções 3.2.2 e 3.2.3, respectivamente. A indisponibilidade desses serviços é crítica para os fluxogramas de operação das aplicações apresentadas na Seção 3.2.3.

Sendo assim, todas as aplicações desenvolvidas nesta tese se beneficiam das características mencionadas, por meio dos serviços disponibilizados, tendo em vista que eles são implantados levando em conta essas particularidades.

3.2.2 O serviço de detecção de seres humanos

O serviço de detecção de seres humanos, proposto neste trabalho, emprega os detectores ACF e ICCF, discutidos na Seção 2.2. Conforme já mencionado, esses detectores fazem parte da família de soluções conhecida como característica de canais filtrados (FCF), que foi durante muitos anos o estado da arte da detecção genérica de pedestres. Essa família de soluções é

conhecida na literatura pelo rápido tempo de detecção e baixa complexidade computacional, principalmente devido ao emprego de árvores de decisão. O esquema rápido de rejeição de exemplos negativos desses classificadores, mencionado na Seção 2.2.3, permite lidar com o fato da maior parte da imagem (fundo) não ser relativa a seres humanos.

De acordo com o que já foi apresentado, neste capítulo a contribuição central se encontra no paradigma utilizado na concepção e nas características agregadas à arquitetura do serviço proposto. A qualidade da detecção, analisada detalhadamente no Capítulo 2, apesar de importante, não é o foco principal. Nesse sentido, o detector ACF é empregado utilizando sua configuração original de parâmetros, tal qual apresentado em (DOLLÁR et al., 2014). Nessa configuração, o ACF é um detector com taxa de processamento aferida próxima a 25 FPS, o que justifica a utilização dele em cascata com o ICCF, conforme será visto adiante. Essa versão apresenta um menor custo computacional que aquelas avaliadas no Capítulo 2 e possui uma LogAvrMR igual a 16.83% (semente aleatória 0).

Além disso, emprega-se também uma versão reduzida do detector ICCF com 2 filtros 5x5 e o conjunto de parâmetros do Modelo 1, que foi apresentado na Seção 2.3. Essa menor quantidade de filtros é importante para tornar o detector ICCF mais rápido na etapa de detecção. Como resultado, esse detector opera em torno de cinco FPS. Além do mais, devido ao procedimento de filtragem apresentado na Seção 2.2, o ICCF possui uma melhor qualidade de detecção que o ACF, com uma LogAvrMR igual a 15.28% (semente aleatória 82). De acordo com o que será visto a seguir, o ICCF é utilizado para filtrar falsos positivos que eventualmente sejam detectados pelo ACF. Isso se deve ao fato do ICCF apresentar uma taxa de FPPI igual a 1,95, ao se considerar todas as detecções por ele retornadas, enquanto o ACF apresenta uma FPPI 2,3 vezes maior, igual a 4,45. A taxas de FPS supracitadas foram aferidas para imagens 640x480. As análises de tempo de processamento e qualidade de detecção foram conduzidas em um computador com processador Intel® Core^{TM} i7-960, 3.20 GHz e 24GB de memória RAM. Os detectores ACF e ICCF foram treinados e testados na base de dados INRIA. É válido ressaltar que o serviço de detecção proposto não utiliza modelos treinados com imagens do espaço inteligente.

Face ao exposto, a escolha dos detectores ACF e ICCF se deve ao fato deles apresentarem, juntos, uma boa relação custo computacional *versus* qualidade de detecção. É importante mencionar também que esses detectores não dependem do uso de GPU. O ACF e o ICCF possuem, somados, um número de parâmetros a serem estimados da ordem de 10^4, quatro ordens de magnitude menor que a VGG16 (aproximadamente 10^8 parâmetros). A VGG16 é uma das soluções baseadas em DCNN mais utilizadas na literatura (SIMONYAN; ZISSERMAN, 2014). Ressalta-se ainda que, mesmo na etapa de detecção, e não somente treinamento, é comum que sejam utilizadas GPU ao se empregar soluções baseadas em DCNN (CAI et al., 2016; ZHANG et al., 2016a). Conforme já mencionado, soluções baseadas em DCNN poderiam ser também utilizadas no fluxograma da solução do serviço proposto. Contudo, um serviço que não depende de *hardware* específico pode ser alocado e distribuído de forma mais flexível nos nós de processamento da infraestrutura, sendo essa uma característica da solução apresentada nesta tese.

Conforme mencionado anteriormente, até mesmo os melhores detectores genéricos possuem o desempenho afetado, ao se analisar a generalização para diferentes bases de dados (ZHANG; BENENSON; SCHIELE, 2017). Isso indica que, quando um detector não é treinado com dados extraídos do ambiente que se está analisando, a acurácia tende a diminuir. Neste caso, a fusão de diferentes metodologias pode ser considerada, de forma a se construir aplicações funcionais para tarefas práticas do mundo real. Para lidar com o desempenho não ideal de detectores, é também importante considerar alguma informação *a priori*. Restrições estruturais ou geométricas do ambiente são boas candidatas. No entanto, deve-se tomar o cuidado de não tornar a solução muito configurável, pois isso implica em uma sobrecarga no estágio de implantação, reduzindo a portabilidade do sistema desenvolvido para diferentes ambientes.

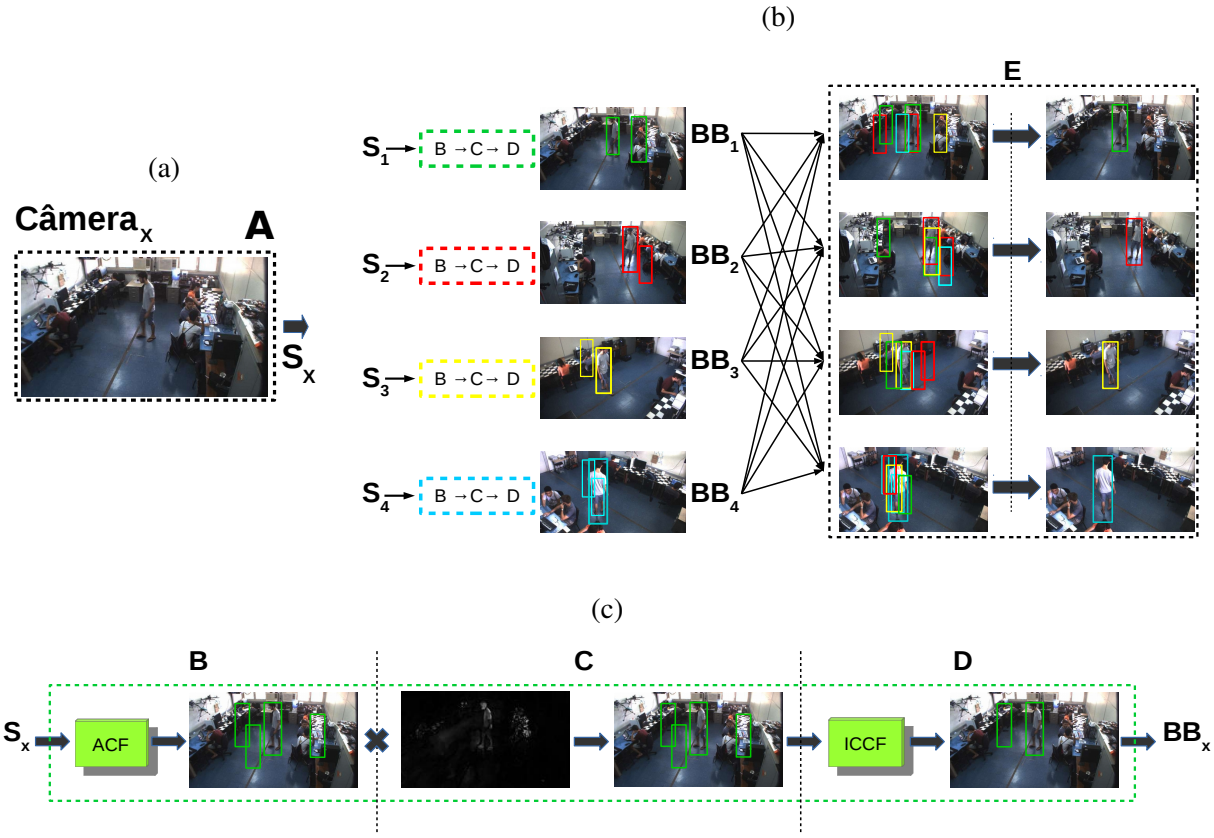
Com isso em mente, a primeira informação *a priori* considerada no serviço de detecção de seres humanos desenvolvido é a calibração das câmeras. Conforme mencionado, esse processo está implementado como um serviço no espaço inteligente. Uma outra consideração é que os seres humanos tocam o plano do chão com seus membros inferiores. Com essas duas restrições, torna-se possível recuperar a posição 3D dos humanos detectados, mesmo empregando apenas uma câmera. No entanto, haja vista que uma rede de câmeras está sendo empregada, os parâmetros de calibração permitem também a estimação da homografia entre as diferentes imagens. Com isso, as janelas (BB) de detecção presentes em uma imagem podem ser reprojetadas em outras imagens do sistema de câmeras. Dessa forma, a correspondência entre diferentes detecções pode ser realizada, de maneira a reduzir o número de falsos positivos. Mesmo com as referidas restrições, o sistema é portável para outros ambientes a um baixo custo de configuração inicial. É importante ressaltar que, se não for necessária a utilização de informação 3D, a calibração das câmeras pode ser evitada.

A Figura 13 mostra o fluxograma de operação do processo de detecção de seres humanos. O serviço proposto (Figura 13b) é composto por dois serviços independentes, que são os serviços de localização de seres humanos e de filtragem de BB. Esses dois serviços são mostrados na Figura 13c e no estágio E da Figura 13b, respectivamente. Em resumo, o serviço de localização emprega os detectores ACF e ICCF, além de um processo de subtração de fundo, de forma a detectar seres humanos nas imagens. Finalmente, essas detecções passam por um processo de refinamento, empregando homografia, por meio do serviço de filtragem de BB. Na sequência, os componentes do serviço de detecção de seres humanos são descritos em detalhes.

No estágio A (Figura 13a), câmeras proveem imagens ($\mathrm{S_x}$, $\mathrm{x} = 1...\mathrm{N}$) por meio de entidades virtuais denominadas *gateways*. No presente trabalho, quatro câmeras são empregadas, logo $\mathrm{N} = 4$. No entanto, um número arbitrário N de *gateways* pode ser instanciado se mais câmeras são disponibilizadas. Esses *gateways* são completamente independentes e podem ser distribuídos entre as estações de trabalho da infraestrutura empregada. Na Figura 13c, do estágio B ao D, o processo de localização de seres humanos ocorre. Cada fluxo de processamento é instanciado como um serviço independente. O serviço de localização de seres humanos toma uma imagem $\mathrm{S_x}$, de uma determinada câmera X, e entrega um conjunto de janelas de detecção $\mathrm{BB_x}$. No estágio B,

o detector ACF é empregado devido a sua alta taxa de FPS e baixa taxa de perda de humanos na imagem (DOLLÁR et al., 2014). No entanto, algumas BB não relativas a seres humanos são também retornadas. De forma a remover essas amostras de detecções que são falsos positivos, um método em cascata é aplicado apenas ao conjunto reduzido de detecções retornadas pelo ACF. Essa escolha faz com que o processo de detecção em cascata não seja computacionalmente proibitivo.

Figura 13 – Fluxograma de operação do processo de detecção de seres humanos: (a) *gateway* da câmera (c) serviço de localização de seres humanos (b) visão geral do serviço de detecção de seres humanos. Em (c), o símbolo $\times$ representa o processo de ponderação dos graus de confiabilidade.



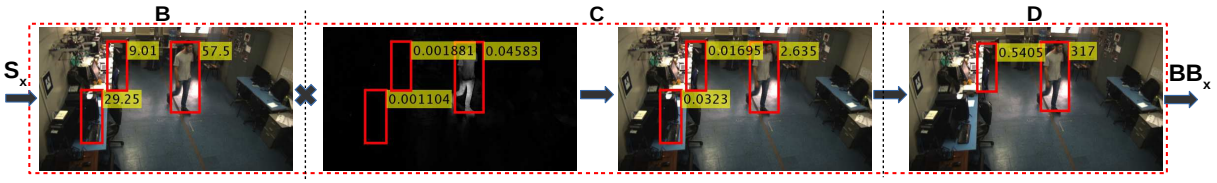
A subtração de fundo (*background subtraction* - BS) é um método comumente empregado na área da visão computacional. No entanto, esse método é muito sensível a variações de iluminação. Por conta disso, ele é utilizado com cautela nesta tese. No estágio C, a subtração de fundo é empregada apenas para ponderar os graus de confiabilidade, relativos às detecções retornadas pelo ACF no estágio B da Figura 13c. O grau de confiabilidade pode ser visto como um índice, que está correlacionado com a probabilidade de uma BB corresponder a um ser humano. Conforme mencionado na Seção 2.2.3, ele é obtido por meio do valor do modelo $F(x)$, descrito na Equação 3. A Equação 9 apresenta o referido processo de ponderação

$$\text{conf}_{NOVO} = \text{conf}_{ANTIGO} \times \text{BB}_{PROB\_BS}, \tag{9}$$

sendo que BB_{PROB_BS} é a soma dos píxeis contidos na BB, divididos pela área da BB, conside-

rando a imagem resultado do processo de subtração de fundo. Por exemplo, dado que os píxeis da imagem com fundo subtraído possuem valores que variam entre 0 e 1, se metade dos píxeis contidos no interior da BB possui valor igual a 1 e a outra metade valor igual a 0, o valor de BB_{PROB_BS} será igual a 0,5. Esse procedimento pode ser melhor compreendido ao se analisar a Figura 14, que é uma versão complementar da Figura 13c. Repare que, nas referidas figuras, o símbolo $\times$ representa o processo de ponderação dos graus de confiabilidade. Os valores, relativos ao termo literal BB_{PROB_BS}, são apresentados na primeira imagem do estágio C. Os graus de confiabilidade das BB, anteriores ao processo de ponderação (conf_{ANTIGO}), são apresentados na imagem relativa ao estágio B da Figura 14. Note que as BB relativas ao estágio B, que estão relacionadas as regiões pretas da imagem com fundo subtraído, tem seus graus de confiabilidade (conf_{NOVO}) reduzidos de maneira mais expressiva. Somente o valor da confiança é modificado e as formas das BB não são alteradas.

Figura 14 – Serviço de localização de seres humanos. O símbolo $\times$ representa o processo de ponderação dos graus de confiabilidade.



Dessa forma, diferentemente de (ALBAWENDI et al., 2015), em que a subtração de fundo é crucial para detectar as posições espaciais de objetos, neste trabalho, ela é utilizada apenas para reduzir o grau de confiabilidade das BB relativas as detecções que são falsos positivos. Eventualmente, variações na imagem com fundo subtraído, causadas por variações anormais de iluminação, podem resultar, no máximo, em um aumento da confiabilidade de falsos positivos. Entretanto, essas eventuais intercorrências podem ser tratadas pelos outros estágios de filtragem do fluxograma do processo de detecção. Essa abordagem faz com que a solução proposta neste trabalho seja mais robusta a variações de iluminação. Após esse processamento, detecções de baixa confiabilidade são descartadas por um limiar de rejeição, empiricamente estimado. A etapa de subtração de fundo, seguida por um limiar de rejeição, pode ser vista como um classificador adicional. Um processo de segmentação semântica, por exemplo, poderia ser utilizado no lugar desse classificador adicional, se o sistema de câmeras não estivesse fixo e a subtração de fundo não pudesse ser utilizada.

No que tange ao modelo do fundo, existem diferentes formas dele ser obtido. Como pequenas variações na imagem com fundo subtraído não são críticas para o processo como um todo, nesta tese emprega-se uma abordagem simples para obter o modelo do fundo. Nos experimentos, o modelo é sempre gerado a partir da média das 10 primeiras imagens capturadas da cena, quando não existe nela qualquer objeto em movimento. Essa abordagem se mostrou suficiente para os casos das aplicações PdC desenvolvidas. No entanto, existem formas mais sofisticadas que podem ser tratadas como trabalhos futuros. Um exemplo seria a criação de um modelo de

fundo, por meio de um conjunto maior de imagens da cena capturado a *priori*. Nesse caso, esse modelo poderia ser carregado no sistema no início da execução de cada aplicação. Uma outra possibilidade é o uso de um modelo adaptativo com um fator de atualização, que agrega menor importância para as imagens antigas da cena.

No estágio D (Figura 13c), outra rodada de remoção de falsos positivos é realizada, empregando o detector ICCF. Conforme visto no Capítulo 2, o ICCF discrimina pedestres do fundo da imagem melhor que o ACF, contudo, a uma menor taxa de FPS. Por isso, o ICCF é utilizado apenas no conjunto reduzido de detecções retornadas pelo estágio C do serviço de localização de seres humanos (Figura 13c). Note que, no estágio D da Figura 14, uma das BB relativas ao estágio C foi eliminada pelo ICCF. O grau de confiabilidade das detecções retornadas pelo estágio D é obtido, então, por meio da multiplicação entre os graus de confiabilidade retornados pelo ICCF e pelo estágio C. Com isso, obtém-se um grau de confiabilidade final mais robusto. Finalmente, as detecções que são geradas pelo estágio D passam ainda por um processo de triagem, no qual detecções de baixa confiabilidade são descartadas.

O último passo do fluxograma de operação do detector de seres humanos é o serviço de filtragem de BB, que é exibido na Figura 13b. Ele emprega homografia para transformar as BB entre as imagens da rede de câmeras do espaço inteligente. Para isso, modela-se a geometria de aquisição da câmera por meio do modelo pinhole, representado na equação seguinte:

$$\lambda_i \tilde{\mathbf{m}}_\mathbf{i} = \mathbf{K_i}\ \mathbf{\Pi}\ [\mathbf{R_i}, \mathbf{T_i}]\ \tilde{\mathbf{M}}, \tag{10}$$

sendo que λ_i é um fator de escala, $\tilde{\mathbf{m}}_\mathbf{i} = [u_i\ v_i\ 1]^T$ é um ponto na imagem da i-ésima câmera, $\mathbf{K_i}$ é a matriz de parâmetros intrínsecos, $\mathbf{\Pi}$ é a matriz de projeção e $[\mathbf{R_i}, \mathbf{T_i}]$ é a matriz de parâmetros extrínsecos composta, respectivamente, por uma matriz de rotação e um vetor de translação. Finalmente, $\tilde{\mathbf{M}} = [x\ y\ z\ 1]^T$ é um ponto 3D no sistema de coordenadas de referência do mundo, no qual as câmeras foram calibradas, e que gera a projeção $\tilde{\mathbf{m}}_\mathbf{i}$ no plano da imagem. Os literais com as acentuações ˜ são representados em coordenadas homogêneas.

É possível reescrever a Equação 10 como

$$\lambda_i \tilde{\mathbf{m}}_\mathbf{i} = \mathbf{A_i}\ [x\ y\ z\ 1]^T, \tag{11}$$

representando o termo $\mathbf{K_i}\ \mathbf{\Pi}\ [\mathbf{R_i}, \mathbf{T_i}]$ por meio da matriz $\mathbf{A_i}$ de dimensão 3×4, cujas colunas são indicadas por $\mathbf{a}_\mathbf{i}^c$. Como se considera que os seres humanos estão sempre tocando o plano do chão, isto é, $z = 0$, as únicas incógnitas da Equação 11 são x, y e λ_i. Note que a matriz $\mathbf{A_i}$ já é conhecida por meio do procedimento de calibração e o ponto $\tilde{\mathbf{m}}_\mathbf{i}$ pode ser obtido a partir da imagem, por exemplo, a localização dos pés dos seres humanos (aproximadamente o ponto médio da base da BB). Consequentemente, as coordenadas x e y, no referencial do mundo, podem ser calculadas utilizando a seguinte equação:

$$\left[\ \mathbf{a}_\mathbf{i}^1\ \middle|\ \mathbf{a}_\mathbf{i}^2\ \middle|\ -\tilde{\mathbf{m}}_\mathbf{i}\ \right] \begin{bmatrix} x \\ y \\ \lambda_i \end{bmatrix} = \left[\ -z\mathbf{a}_\mathbf{i}^3 - \mathbf{a}_\mathbf{i}^4\ \right]. \tag{12}$$

É dessa forma que se obtém a posição 3D dos seres humanos detectados. Uma vez que a informação 3D $[x, y, z]$ é conhecida, pode-se usar a homografia $\mathbf{H}_{\mathbf{j}}^{\mathbf{i}}$ entre as i-ésima e j-ésima câmeras, de forma a realizar a correspondência entre os pontos $\tilde{\mathbf{m}}_{\mathbf{i}}$ e $\tilde{\mathbf{m}}_{\mathbf{j}}$ de suas respectivas imagens, conforme apresentado na equação seguinte:

$$\tilde{\mathbf{m}}_{\mathbf{i}} = \mathbf{H}_{\mathbf{j}}^{\mathbf{i}} \, \tilde{\mathbf{m}}_{\mathbf{j}}. \tag{13}$$

Usando a homografia, é possível então reprojetar o ponto médio da base da BB de uma imagem para qualquer outra da rede de câmeras. A altura e a largura das BB são ajustadas, na reprojeção da BB da imagem da câmera j para a da câmera i, por meio da equação seguinte:

$$\begin{bmatrix} w_{ni} \\ h_{ni} \end{bmatrix} = \frac{\lambda_j}{\lambda_i} \begin{bmatrix} w_{nj} \\ h_{nj} \end{bmatrix}, \tag{14}$$

em que w_{nx} e h_{nx} são, respectivamente, a altura e a largura da n-ésima BB, relativa a imagem da x-ésima câmera. Repare que os fatores de escala, relativos às câmeras i e j, são empregados.

Se uma BB presente em uma imagem não possui pelo menos uma correspondente em qualquer outra imagem, ela é descartada. De forma a comparar BB, utiliza-se a métrica IoU, apresentada na Seção 2.3.1. Uma correspondência é considerada se a IoU é maior que 0,5. A Figura 15 demonstra o procedimento de filtragem de BB. Repare que a BB relativa ao falso positivo, na imagem j, é eliminada, pois nenhuma BB da imagem i foi projetada na imagem j. As demais BB são mantidas, pois possuem correspondentes, advindas de outra imagem, com IoU maior que 0,5. É importante mencionar também que, se um ser humano for detectado em apenas uma câmera, ele também terá sua BB descartada, contribuindo para um aumento da MR do serviço de detecção.

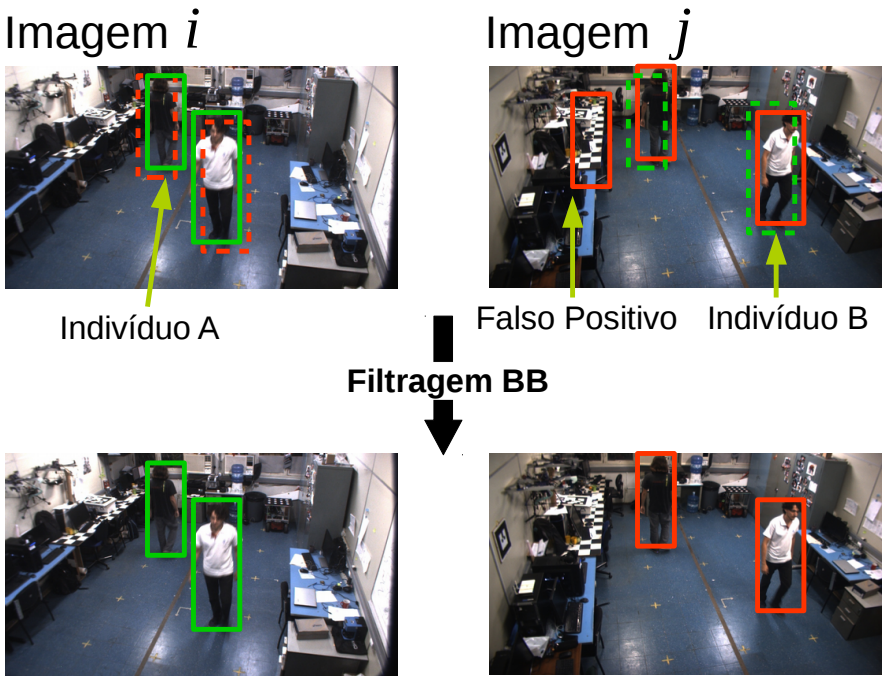
É importante mencionar que todas as imagens empregadas no detector de seres humanos, nos intervalos entre amostragens, são sincronizadas. Isso é feito de forma a se garantir que, para imagens com intercessão, a mesma cena esteja sendo capturada, em um mesmo instante de tempo. O sincronismo entre as câmeras é disponibilizado por um dos serviços de suporte mencionados na Seção 3.2.1.3.

No final do fluxograma de operação do detector de seres humanos, espera-se que apenas aqueles que estejam em pé sejam finalmente detectados nas imagens. No escopo deste trabalho, apenas seres humanos nessa situação são considerados como aqueles que demandam serviços do robô presente no ambiente. É importante mencionar também que o serviço de detecção de seres humanos pode funcionar mesmo se apenas uma câmera estiver provendo imagens. No entanto, neste caso, o fluxograma de operação perde a robustez e os benefícios obtidos por meio da rede de câmeras, conforme verificado nos experimentos.

3.2.3 Aplicações de espaços inteligentes

Nesta tese, três aplicações provas de conceito (PdC), inseridas no contexto de espaços inteligentes, são desenvolvidas de forma a demonstrar a eficácia do serviço de detecção de seres

Figura 15 – Eliminação de falsos positivos empregando homografia. De forma a simplificar o entendimento, apenas duas imagens são empregadas. A cor verde representa a imagem i, enquanto a cor vermelha representa a imagem j. As BB pontilhadas foram transformadas de uma imagem de origem para uma de destino.



humanos proposto. A Figura 16 ilustra as tarefas executadas em cada aplicação desenvolvida.

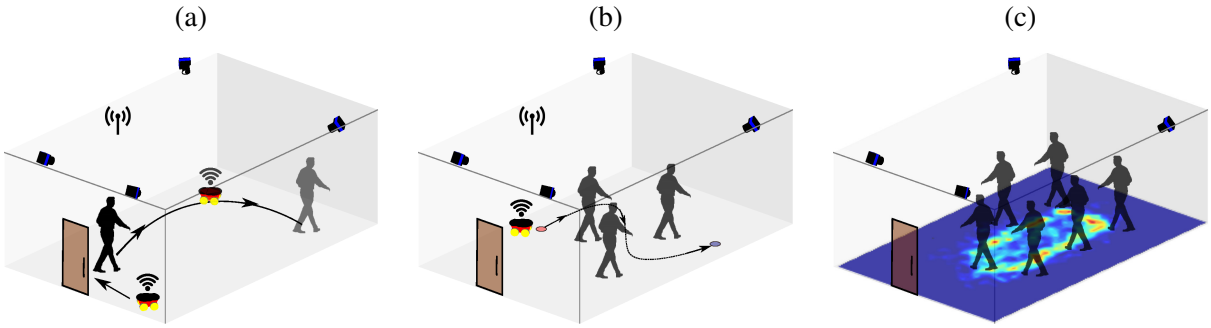
A primeira aplicação é uma tarefa de seguimento de seres humanos, executada por um robô (Figura 16a). O indivíduo de interesse é aquele que acaba de entrar em uma sala. Essa é uma estratégia comum, pois ao acessarem um determinado ambiente, usuários podem estar necessitando de alguma assistência.

A Figura 16b exibe a segunda tarefa, na qual um robô tem que navegar pelo espaço inteligente desviando dos seres humanos nele presentes. Essa aplicação é importante, por exemplo, para complementar a primeira tarefa, de forma que o robô siga um indivíduo sem colidir com os demais presentes no ambiente. Essas aplicações são PdC de diferentes tarefas do dia a dia e podem ser executadas em ambientes em que uma rede de câmeras encontra-se disponível, tal como bancos, museus, *shoppings* e praças públicas.

A Figura 16c mostra o mapa de ocupação cumulativo do espaço inteligente. Esse mapa é dito cumulativo, pois tem o objetivo de consolidar os locais mais visitados em um ambiente ao longo do tempo. Dessa forma, uma análise temporal é realizada. Esse tipo de aplicação é útil, por exemplo, para determinar dinamicamente os melhores locais para se posicionar placas de publicidade ou mesmo determinar o preço de venda ou aluguel de salas em um *shopping*. Os locais mais visitados são aqueles nos quais os produtos terão uma maior visibilidade por parte dos consumidores. Nesse sentido, espera-se que salas posicionadas próximo a esses locais possuam uma maior valorização.

A Figura 17 mostra os serviços específicos relacionados a visão computacional, que compõem

Figura 16 – Aplicações de: (a) seguimento e (b) desvio de seres humanos e criação de (c) mapa de ocupação cumulativo.



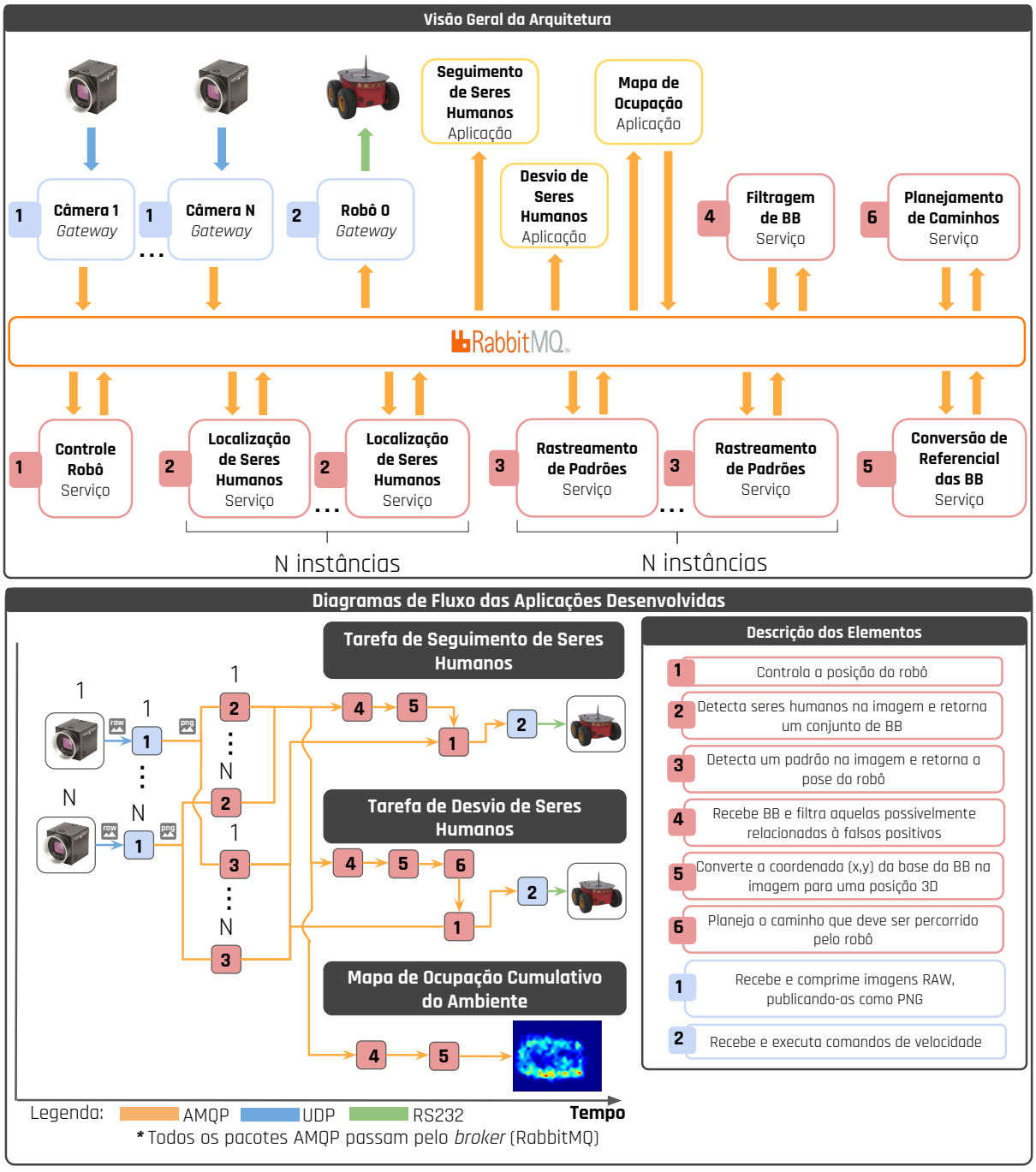
as aplicações. Esses serviços e seus respectivos inter-relacionamentos, dentro da arquitetura do espaço inteligente, foram abordados na Seção 3.2.1. As Figuras 17 e 12 são complementares. A primeira ilustra o fluxo de mensagens entre os serviços e aplicações, enquanto a última apresenta a distribuição dessas entidades virtuais em diferentes camadas da arquitetura. Serviços de suporte, tais como o de calibração e sincronismo, não são mostrados no diagrama da Figura 17, de forma a proporcionar uma melhor visualização.

Na Figura 17, o serviço de Rastreamento de Padrões é responsável por recuperar e publicar a posição 3D do robô, usando um padrão geométrico reconhecido por meio das imagens da rede de câmeras. O método utilizado para detectar o padrão geométrico é derivado da abordagem empregada em (RAMPINELLI et al., 2014). Por meio dessa odometria visual, o robô pode ser controlado por um serviço denominado Controle Robô. Os serviços de Localização de Seres Humanos e Rastreamento de Padrões consomem os quadros publicados pelos *gateways* das câmeras. O serviço de Localização de Seres Humanos provê BB para o serviço de Filtragem de BB. Por sua vez, o serviço de Conversão de Referencial das BB é responsável por projetar, no referencial 3D do mundo, BB publicadas pelo serviço de Filtragem de BB. Uma vez que as informações tridimensionais relativas aos seres humanos e ao robô estejam disponíveis, as aplicações podem ser executadas.

A única aplicação que não emprega o serviço de Rastreamento de Padrões é a relativa ao mapa de ocupação cumulativo. É importante notar que todas as aplicações mostradas na Figura 17 são independentes e podem ser executadas ao mesmo tempo no espaço inteligente. Cada fluxo da Figura 17 é apenas uma ocorrência do laço principal de cada aplicação. Os fluxos são sistematicamente repetidos durante as execuções das aplicações.

Os fluxos originados a partir dos serviços de Localização de Seres Humanos e Rastreamento de Padrões são assíncronos e executados em paralelo. Nas aplicações de seguimento e desvio de seres humanos, esses fluxos convergem para o serviço de controle do robô, tendo em vista que eles proveem as informações necessárias para executar as aplicações propostas. As N instâncias do serviço de Localização de Seres Humanos e Rastreamento de Padrões são também executadas em paralelo e isso mostra o paralelismo intrínseco ao paradigma da arquitetura empregada. Essa característica é de particular importância, devido ao fato de detectores de seres humanos e

Figura 17 – Inter-relacionamento entre serviços e aplicações.



objetos serem, normalmente, tarefas que demandam o emprego de uma quantidade considerável de processamento computacional. Nesse caso, esses processos podem ser distribuídos sobre a infraestrutura do espaço inteligente, durante os estágios de teste, de forma a utilizar aqueles nós com maiores quantidades de recursos computacionais disponíveis.

3.3 Experimentos

Nesta seção, experimentos são conduzidos considerando o serviço de detecção de seres humanos proposto, no contexto das aplicações desenvolvidas para o espaço inteligente. A ideia é analisar as principais características da solução proposta, por meio da observação do comportamento, de cada aplicação, durante a operação em tempo real.

Os seguintes pontos principais da solução proposta são analisados durante os experimentos:

- A habilidade do serviço proposto em atender, de forma flexível, a diferentes aplicações em um espaço inteligente;

- A efetividade do serviço de detecção de seres humanos em atender as demandas das aplicações desenvolvidas, quando atuando em um espaço inteligente baseado em uma rede de câmeras. Além da qualidade da detecção e localização 3D, demonstra-se o correto funcionamento das aplicações, ao mesmo tempo que se provê características como paralelismo, confiabilidade e escalabilidade;

- O atendimento aos requisitos de tempo real das aplicações com interação homem-máquina[8], devido à cooperação harmoniosa entre o serviço proposto e aqueles existentes na plataforma do espaço inteligente.

Ao longo dos experimentos, avalia-se também a capacidade dos detectores genéricos de seres humanos ACF e LDCF em atender, isoladamente, as aplicações em termos de qualidade de detecção. Uma comparação com o serviço de detecção proposto é realizada. A análise comparativa se limita aos detectores ACF e LDCF, por eles fazerem parte do escopo de soluções possíveis de serem empregadas, dentro do contexto do espaço inteligente utilizado.

É importante mencionar que os detectores ACF e LDCF, na referida análise comparativa, não substituem o ICCF no fluxo de operação da Figura 13c. Eles são analisados de forma isolada em cada câmera. A ideia é justamente demonstrar a importância do esquema em cascata, apresentado na Figura 13, para o correto funcionamento das aplicações. Por meio dessa análise, pode-se fazer um paralelo entre soluções aplicadas a situações controladas (Capítulo 2) e problemas reais executados em tempo real.

Por fim, é válido ressaltar ainda que, por meio da análise de qualidade realizada, não se deseja reivindicar um método de detecção, baseado em um sistema de câmeras, estado da arte. No tocante ao funcionamento das aplicações, a análise de qualidade é uma forma, indireta, de validar as características do serviço proposto.

[8] É importante lembrar que o termo interação homem-máquina, neste trabalho, se refere a percepção da localização 3D dos seres humanos, presentes no ambiente, ao mesmo tempo que se realiza a navegação do robô.

3.3.1 Materiais e métodos

3.3.1.1 Infraestrutura

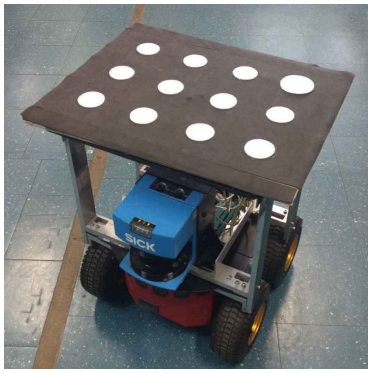
A Tabela 9 descreve os elementos presentes na infraestrutura de *hardware* do espaço inteligente. As quatro câmeras são os únicos sensores empregados. Conforme mencionado anteriormente, o robô é rastreado usando um serviço de odometria visual, que detecta um padrão que se encontra anexado ao robô, conforme exibido na Figura 18. O odômetro interno ao robô e o sensor laser, que é apresentado na Figura 18, não são utilizados.

Tabela 9 – Infraestrutura do espaço inteligente.

Dispositivo	Descrição Processador/Memória RAM
Nó 1	i7-6850K/128 GB
Nó 2	i5-3570/16 GB
Nó 3	i5-4460S/8 GB
Nó 4	E5504/4 GB
Câmeras 1...4	Blackfly BFLY-PGE-09S2C
Robô	Pioneer P3-AT

Pode-se notar, a partir da Tabela 9, que o Nó 1 possui uma quantidade relativamente alta de memória RAM. No entanto, as quatro instâncias do serviço de Localização de Seres Humanos, que é o serviço de maior custo computacional, utilizam apenas 800 MB de memória RAM juntas. Exceto pelo robô, os demais dispositivos possuem interfaces de rede Gigabit Ethernet.

Figura 18 – Robô utilizado nos experimentos.



As câmeras utilizadas capturam imagens na resolução 1288x728, no entanto, no processo de detecção as imagens são subamostradas para a resolução 512x291. Isso é feito para reduzir o tempo demandado pelo processo de detecção. As câmeras estão dispostas no espaço inteligente de acordo com o apresentado na Figura 19. Amostras de imagens de cada uma das câmeras são apresentadas na Figura 20. A área destacada na Figura 19 representa a área de operação do espaço inteligente, onde a intercessão entre as imagens das câmeras permite o funcionamento do processo de detecção, conforme será discutido na Seção 3.3.2.3. O espaço inteligente possui uma área de 4,8 m x 7,3 m, com uma área de operação aproximada de 4,8 m x 5,5 m. O ambiente não

possui coloração homogênea e, embora o plano do chão apresente uma cor característica, podem ser observados pontos de saturação devido à iluminação artificial. Por fim, existem no espaço de trabalho diferentes objetos, o que torna a cena rica em informação e dificulta o processo de detecção.

Figura 19 – Posicionamento das câmeras utilizadas e espaço de trabalho.

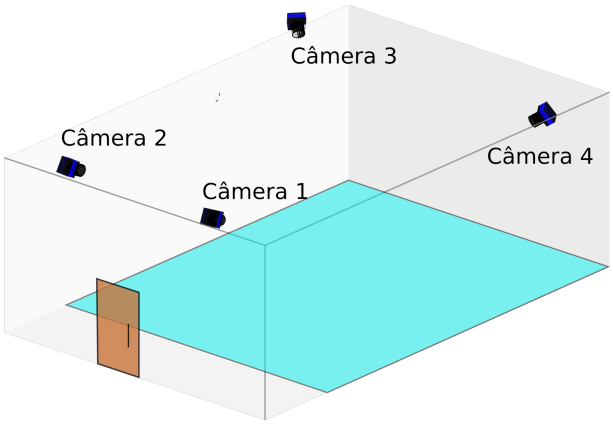
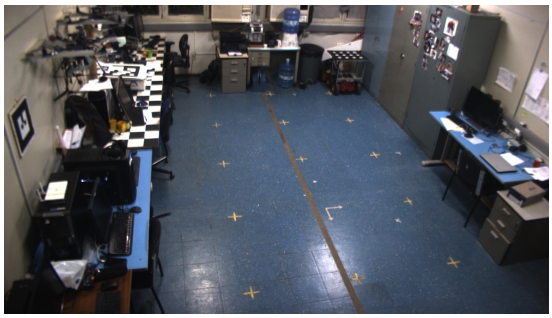


Figura 20 – Amostras de imagens das câmeras utilizadas.

(a) Câmera 1.

(b) Câmera 2.



(c) Câmera 3.

(d) Câmera 4.

3.3.1.2 Detecção de pedestres

Os detectores de pedestre ACF e ICCF, empregados no fluxograma de operação do serviço de detecção, foram treinados utilizando a base de dados INRIA, conforme descrito no Capítulo 2. Nenhuma imagem do espaço inteligente foi utilizada no treinamento desses detectores. O mesmo se aplica ao detector LDCF, empregado nas comparações realizadas nos experimentos. Conforme

mencionado anteriormente, o ICCF utiliza o conjunto de parâmetros do Modelo 1, enquanto os detectores ACF e LDCF utilizam suas configurações originais de parâmetros, apresentadas em (DOLLÁR et al., 2014) e (NAM; DOLLÁR; HAN, 2014), respectivamente. Esses detectores não são treinados especificamente para lidar com o problema de oclusão severa e, por isso, a redundância provida pelo sistema de câmeras é importante. Adicionalmente à análise qualitativa realizada, são disponibilizados dados quantitativos acerca do serviço de detecção de seres humanos, de forma a avaliar seus pontos fortes e fracos.

Com respeito à avaliação quantitativa no plano da imagem, a taxa de perda (MR), o número de falsos positivos por imagem (FPPI) e a precisão (PR) são as métricas empregadas. Conforme já mencionado no Capítulo 2, a MR é a fração total de seres humanos não identificados e FPPI é o número de detecções que são falsos positivos dividido pelo número de imagens do experimento avaliado. A precisão representa a fração de detecções positivas em relação ao total de detecções retornadas pelo serviço de detecção. É importante mencionar, novamente, que uma detecção positiva é considerada se existe uma IoU maior que 0,5 entre a detecção avaliada e uma anotação presente na base de dados. Novamente, uma anotação pode ser correspondida apenas uma vez.

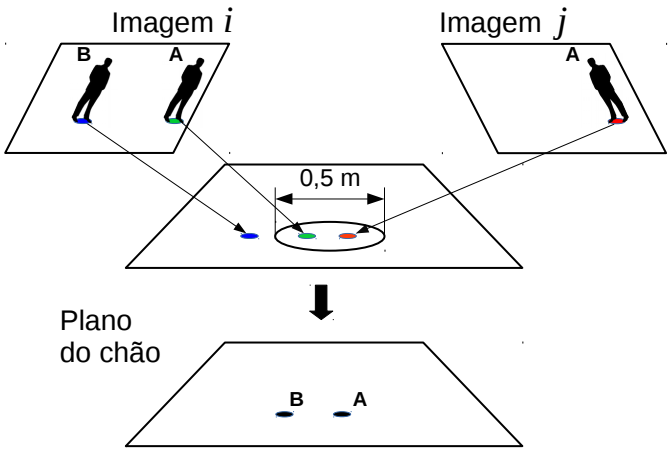
No que diz respeito à avaliação no plano do chão, o número de verdadeiros positivos (*true positives* - TP), falsos positivos (FP) e falsos negativos (FN) são apresentados. Além disso, o erro de localização entre as detecções positivas e suas anotações correspondentes, na base de dados, é analisado. De forma a considerar uma correspondência entre um ponto que representa uma detecção e uma dada anotação, ambos no plano do chão, uma distância euclidiana menor que 0,5 m, entre eles, é demandada[9]. Novamente, uma anotação pode ser correspondida apenas uma vez.

De forma a computar os valores das métricas supracitadas, as imagens do espaço inteligente, empregadas nos experimentos, tiveram todos os humanos nelas presentes anotados por meio de BB. Para calcular as métricas relacionadas ao plano do chão, as anotações foram obtidas por meio da projeção, nesse plano, das BB anotadas manualmente nas imagens (detector perfeito). A Figura 21 exibe o referido processo. Repare que, quando o mesmo ser humano é anotado em mais de uma imagem (humano A), ele irá gerar dois pontos no plano do chão. Nesse caso, esses pontos são agrupados em um único ponto, se estão dentro de um círculo de diâmetro igual a 0,5 m. Esse ponto é considerado a anotação, no plano do chão, relativo ao humano A. Em uma análise complementar (Seção 3.3.2.4), calcula-se também o erro de localização em relação a pontos específicos e uniformemente distribuídos em relação ao plano do chão do espaço inteligente. Nesses casos, as anotações foram obtidas por meio de medições diretas realizadas na superfície do chão do laboratório empregado nos experimentos.

Finalmente, sem perda da autocompletude da presente tese de doutorado, um código-fonte de amostra do serviço de detecção de seres humanos e todos os dados anotados estão disponíveis publicamente em (ALMONFREY et al., 2018b).

[9] Esse valor é normalmente considerado o diâmetro médio da área projetada pelo corpo humano, em pé, no plano do chão.

Figura 21 – Processo de agrupamento de pontos. De forma a simplificar o entendimento, apenas duas imagens são empregadas.

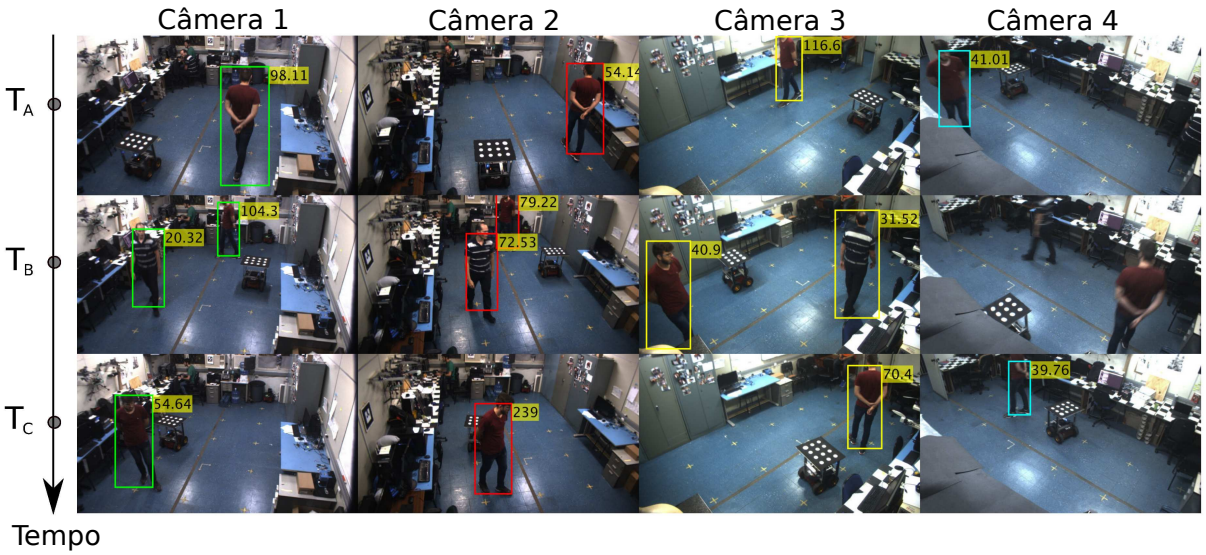


3.3.2 Resultados Experimentais

3.3.2.1 Tarefa de seguimento de seres humanos

Neste experimento, o serviço de detecção proposto é utilizado para gerar pontos de referência para o sistema de controle do robô, durante a tarefa de seguimento de seres humanos. A Figura 22 mostra o processo de detecção em três diferentes instantes de tempo (T_A, T_B e T_C), ao longo da navegação do robô. Além disso, os números mostrados nas imagens são os graus de confiabilidade de cada detecção.

Figura 22 – Detecção durante a tarefa de seguimento de seres humanos.



Repare que, na Figura 22, no instante T_B, outro indivíduo está presente no espaço de trabalho, de forma a mostrar que o sistema pode detectar mais de um ser humano ao mesmo tempo, ao longo dos experimentos. No entanto, o robô está configurado para continuar seguindo o primeiro humano atendido pelo serviço. Apesar do fato de, em alguns instantes, um dado indivíduo não

ser detectado por uma determinada câmera, ele é detectado pela maioria delas na maior parte do tempo. Desta forma, o emprego de uma rede de câmeras aumenta a taxa de detecção de seres humanos. É importante mencionar que alguns erros são esperados no processo de detecção, em algumas imagens, tendo em vista que o detector não foi treinado com imagens do espaço inteligente.

Por completude, a Tabela 10 exibe uma análise, no plano da imagem, da acurácia do serviço de detecção de seres humanos. A ideia dessa análise não é reivindicar detecção ao nível do estado da arte, tendo em vista que esse não é o foco do presente trabalho. Deseja-se avaliar apenas a qualidade da detecção, no contexto dos requisitos da aplicação. Na referida tabela, o desempenho dos detectores ACF e LDCF são também apresentados. Ao comparar o serviço de detecção de seres humanos com os outros dois detectores da literatura, pretende-se demonstrar a importância do fluxograma de operação completo para o sucesso da aplicação. Repare que a mínima estruturação imposta ao espaço físico de trabalho foi importante para se ter sucesso em situações nas quais detectores genéricos de seres humanos falham.

A partir da Tabela 10, é possível notar que a PR geral (em negrito) do serviço de detecção de seres humanos é consistentemente maior que as apresentadas pelos demais detectores. No entanto, o método proposto neste trabalho tem maior MR, o que é compensado pelo fato de se usar uma rede de câmeras. A coluna relacionada a MR* mostra que a MR geral (em negrito), do serviço de detecção, diminui quando considerando informação advindas de todas as câmeras. Essa menor MR indica que humanos perdidos em uma dada câmera são detectados em alguma outra, conforme considerado anteriormente. Nesse sentido, um bom compromisso entre precisão e perda de detecções é obtido.

Sem o procedimento em cascata do processo de detecção, para remoção de FP, e o emprego de uma rede de câmeras, para diminuir a MR, os detectores ACF e LDCF não seriam capazes de executar a tarefa proposta. É importante mencionar que, mesmo se o ACF e o LDCF estivessem empregando uma rede de câmeras, as precisões deles seriam inapropriadas para a aplicação, tendo em vista que eles não possuem um fluxograma de operação para eliminar FP. A Figura 23 ilustra o maior número de FPPI do ACF e LDCF quando comparados ao método proposto neste trabalho.

O limiar estabelecido para a IoU (lim_{IOU}), de forma a considerar uma correspondência entre uma detecção e uma anotação, foi reduzido para 0,3 somente para a análise da acurácia das BB reprojetadas. Essa forma é mais justa para ilustrar a acurácia, neste caso, tendo em vista que as detecções estão aproximadamente tocando o plano do chão. Desta forma, algumas variações acerca da localização das BB reprojetadas podem ser observadas após a transformação por meio da homografia. No entanto, conforme mencionado anteriormente, durante os experimentos, um limiar de IoU igual a 0,5 foi empregado. Os seres humanos localizados nas bordas das imagens não são levados em conta na análise da acurácia. Embora casos com oclusões não sejam o foco principal, no cenário razoável de avaliação da área da detecção de pedestres, conforme mencionado na Seção 2.3.1, seres humanos sobre efeito de oclusão são considerados na análise

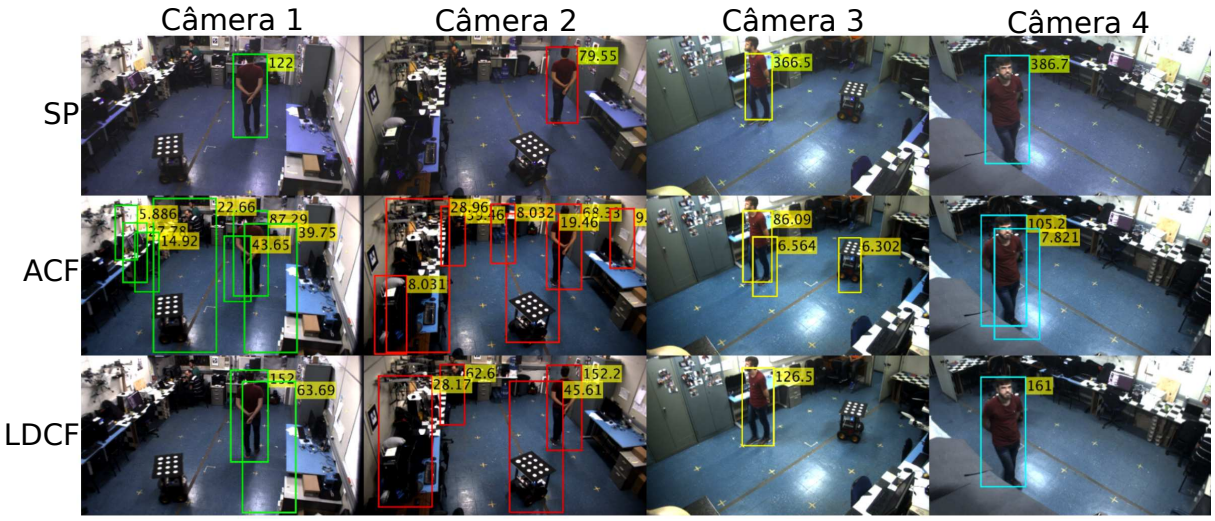
realizada neste capítulo.

Tabela 10 – Análise da detecção de seres humanos no plano da imagem para a primeira aplicação, considerando o serviço proposto e os detectores ACF e LDCF.

Dispositivos	Serviço proposto				ACF			LDCF		
	MR (%)	MR* (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)
Câmera 1	$14,0$	$11,2$	$5,4 \cdot 10^{-3}$	$99,1$	$4,2$	$4,50$	$14,3$	$7,0$	$0,81$	$47,3$
Câmera 2	$26,4$	$11,2$	$2,7 \cdot 10^{-2}$	$94,9$	$4,8$	$6,40$	$9,2$	$5,6$	$2,70$	$19,1$
Câmera 3	$38,7$	$6,6$	$4,9 \cdot 10^{-2}$	$86,8$	$7,5$	$0,92$	$36,7$	$7,5$	$0,14$	$79,0$
Câmera 4	$62,3$	$16,4$	$3,8 \cdot 10^{-2}$	$71,4$	$24,6$	$0,27$	$48,4$	$21,0$	$3,8 \cdot 10^{-2}$	$86,5$
Todas as Cameras	$30,3$	$\mathbf{10,8}$	$3,0 \cdot 10^{-2}$	$\mathbf{93,1}$	$8,0$	$3,02$	$15,3$	$8,7$	$0,92$	$36,9$

*As BB de imagens de outras câmeras são projetadas na imagem da câmera analisada. Isso mostra a vantagem do emprego de uma rede de câmeras.

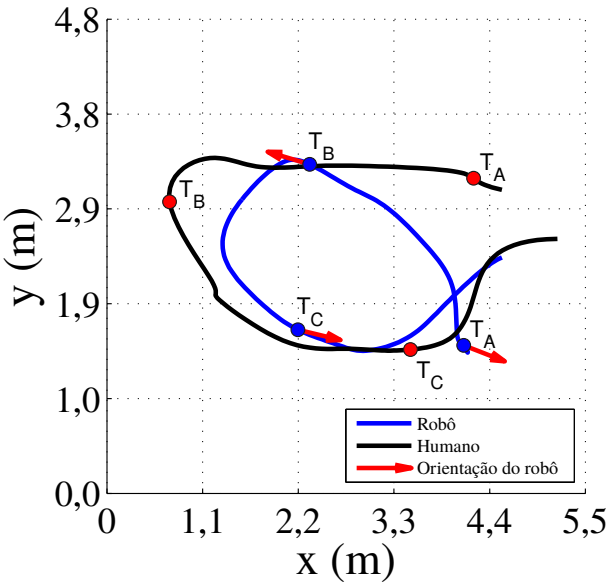
Figura 23 – Falsos positivos relativos ao serviço proposto (SP - primeira linha), ACF (segunda linha) e LDCF (terceira linha).



O Gráfico 6 mostra as trajetórias descritas pelo ser humano e o robô durante o experimento. É possível notar, a partir da mesma figura, que o serviço de detecção de seres humanos provê informação adequada para a tarefa que está sendo executada. Os instantes $\mathrm{T_A}$, $\mathrm{T_B}$ e $\mathrm{T_C}$, mostrados na Figura 22, são destacados na trajetória. Somente a trajetória do ser humano que está sendo seguido pelo robô é exibida para uma melhor visualização. No início do experimento (instante T_A), o robô encontra-se orientado 330^o no sentido anti-horário. Neste caso, ele tem que rotacionar aproximadamente 120^o, em seu próprio eixo, de forma a iniciar a tarefa de seguimento.

Conforme mencionado anteriormente, as posições tridimensionais são obtidas a partir das BB nas imagens capturadas durante uma mesma janela de amostragem. Desta forma, para qualquer humano, múltiplas posições podem ser obtidas, tendo em vista que cada câmera provê uma BB detectada, a partir das quais as posições tridimensionais são obtidas. Assim como foi feito para as anotações na Figura 21, em todos os experimentos deste trabalho, as posições dos seres humanos no plano do chão, obtidas por meio das BB detectadas, são agrupadas se estão internas a uma região de diâmetro igual a 0,5 m. Desta forma, obtém-se a localização dos seres humanos no espaço 3D. Durante a navegação do robô, de forma a aumentar a robustez contra perdas pontuais

Gráfico 6 – Trajetórias durante a tarefa de seguimento de seres humanos.



de detecção, uma memória é utilizada para acumular as últimas 10 posições do ser humano que está sendo seguido pelo robô. O valor médio dessas medições é considerada a posição do indivíduo.

A Tabela 11 apresenta as métricas avaliadas também no plano do chão, para o fluxograma do processo de detecção. Note que a MR no plano do chão $(\frac{\mathrm{FN}}{\mathrm{NGTS}})$ é aproximadamente igual a 13%, que está próximo ao valor geral da MR* ao se considerar informações advindas de todas as câmeras (Tabela 10). Esse baixo valor de MR é ainda mais atenuado pela memória empregada durante a navegação do robô, mas que não é considerada na avaliação de qualidade deste experimento. Esse procedimento ajuda à tarefa de controle em situações transientes, quando um ser humano é perdido. Note também que o baixo número de FP está de acordo com o baixo número de FPPI computados no plano da imagem. É importante mencionar também que a avaliação no plano do chão está correlacionada com a avaliação no plano da imagem, porém, elas não estão relacionadas diretamente de forma numérica. Ainda, a partir da Tabela 11, pode-se verificar que o sistema possui um erro de localização (LOCERR) menor que um quarto do diâmetro médio da área do corpo de um ser humano, em pé, projetada no plano do chão (0,5 m).

Tabela 11 – Análise, no plano do chão, do processo de detecção ao longo da primeira aplicação.

TP	FP	FN	NDTS	NGTS	LOCERR (m)
129	5	20	134	149	$0,11$

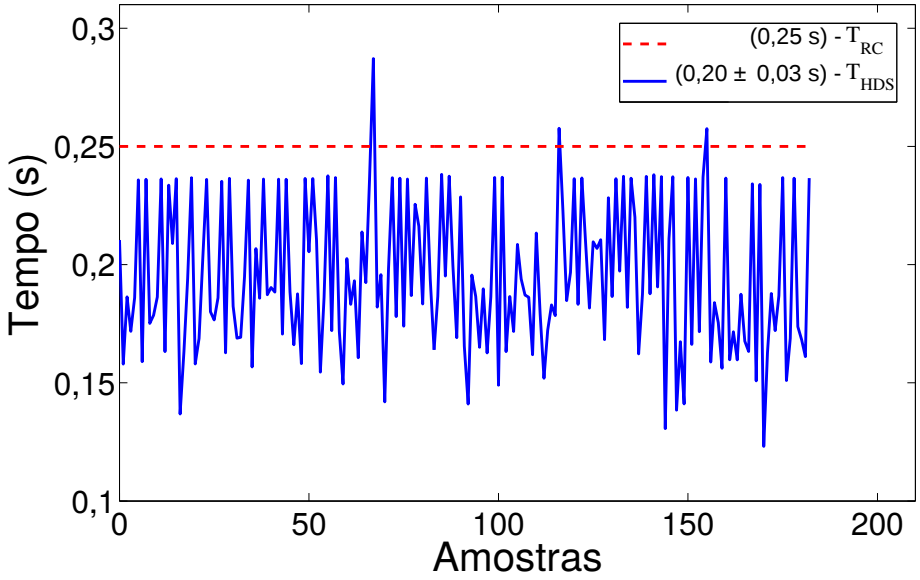
NDTS - Representa o número total de detecções;
NGTS - Esse é o número de anotações estimadas, no plano do chão, empregando as BB anotadas no plano da imagem;
LOCERR - Representa o erro médio de localização entre uma detecção positiva e sua anotação correspondente.

Consequentemente, como pode ser visto, o serviço de detecção de seres humanos proposto apresenta um baixo número absoluto de FP, que atende as demandas da aplicação proposta, ao mesmo tempo que mantêm uma MR apropriada. Novamente, esse baixo valor de FP é alcançado graças ao encadeamento em cascata, no fluxograma de operação do serviço, de soluções que

eliminam detecções que são falsos positivos. Ao mesmo tempo que provê características como escalabilidade, paralelismo e confiabilidade, o serviço proposto atende aos requisitos necessários para o correto funcionamento da aplicação.

Finalmente, o Gráfico 7 apresenta uma análise de tempo de reposta do serviço de detecção de seres humanos. Um tempo $\mathrm{T_{HDS}}$ maior que $\mathrm{T_{RC}}$ significa que o serviço de detecção não está sendo capaz de processar as imagens em um intervalo de tempo adequado, demandado pelo sistema de controle do robô. Um tempo $\mathrm{T_{HDS}}$ menor ou igual a $\mathrm{T_{RC}}$ implica que o tempo de detecção do serviço proposto está em acordo com os requisitos de tempo da aplicação. Na prática, espera-se que o valor de $\mathrm{T_{HDS}}$ oscile em torno de um valor ligeiramente menor que o de $\mathrm{T_{RC}}$.

Gráfico 7 – Análise de tempo de resposta do serviço proposto durante o primeiro experimento.



$\mathrm{T_{RC}}$ - é o máximo intervalo de tempo no qual o serviço de controle do robô pode receber a informação de posição, dos indivíduos presentes no ambiente, e ainda operar o robô adequadamente. Esse valor é definido pelo intervalo de amostragem da câmera, que opera a 4 FPS;
$\mathrm{T_{HDS}}$ - é o tempo médio aproximado medido entre a captura de uma imagem e a entrega de comandos de controle por parte do robô. Esse é então o tempo de processamento do serviço de detecção proposto. Esse tempo foi obtido por meio da média de 180 amostras, sendo que o desvio padrão também é apresentado.

Para $\mathrm{T_{HDS}}$, os valores exibidos, na legenda do Gráfico 7, representam a média e o desvio padrão das amostras das medições realizadas ao longo do tempo. O valor de $\mathrm{T_{RC}}$, também exibido na legenda, é o máximo intervalo de tempo desejado, no qual o serviço de controle do robô deve receber as informações de posição dos seres humanos. Conforme indicado pelo gráfico de trajetória e confirmado pela análise do tempo de resposta do serviço de detecção, o método proposto foi capaz de servir à aplicação, ao mesmo tempo que atende aos seus requisitos de tempo. Na média, o serviço de detecção é capaz de realizar todo o trabalho demandado dentro de um intervalo de amostragem da câmera, não inserindo, portanto, atrasos sensíveis aos usuários.

O sistema de controle do robô, idealmente, tem que possuir um laço de operação com período próximo ao intervalo de amostragem da câmera. Nos experimentos deste capítulo, a câmera foi ajustada para quatro FPS, apresentando dessa forma um período de amostragem igual a $0,25$ s, conforme $\mathrm{T_{RC}}$. Esse valor se justifica como uma questão de projeto, pois esse

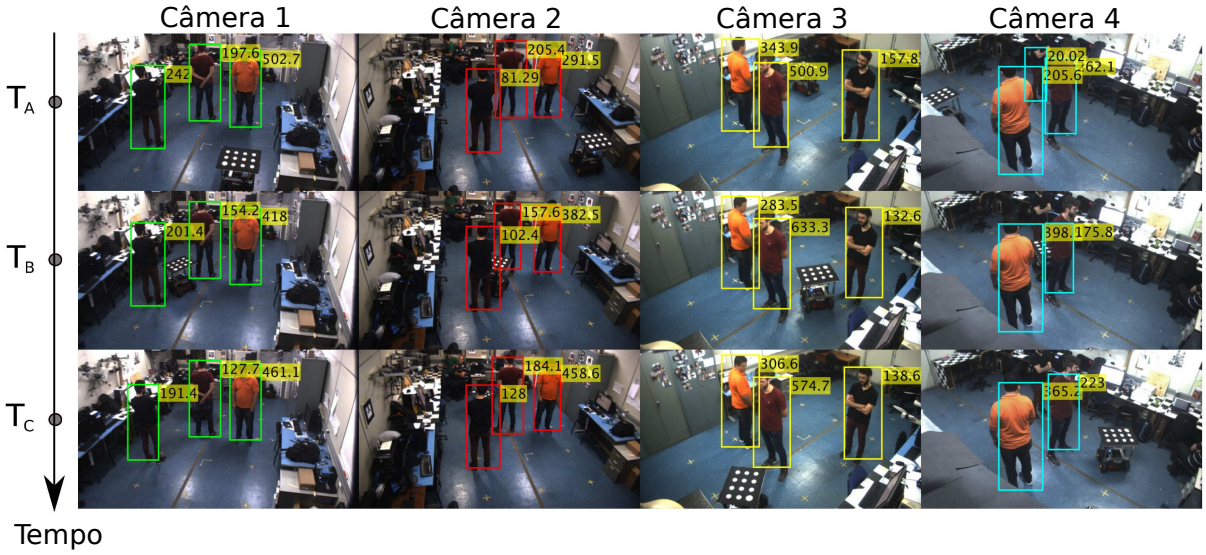
intervalo de amostragem é adequado para realizar as operações, relacionadas à integração de diferentes tecnologias, existentes no âmbito do espaço inteligente. Essa taxa de amostragem é também compatível com as velocidades de deslocamento dos seres humanos e robôs no espaço de trabalho.

De forma adicional aos experimentos apresentados, nesta tese, relativos à tarefa de seguimento de seres humanos, um conjunto de vídeos acerca do presente experimento é disponibilizado em (ALMONFREY, 2018).

3.3.2.2 Tarefa de desvio de seres humanos

Neste experimento, o sistema de controle do robô emprega o serviço de detecção proposto, de forma a executar uma tarefa de desvio de seres humanos ao longo da navegação. Os seres humanos presentes no ambiente são tratados como obstáculos e a trajetória a ser executada pelo robô é planejada por meio de uma estratégia de planejamento de caminhos apresentada em (ŞUCAN; MOLL; KAVRAKI, 2012). A Figura 24 mostra o processo de detecção de seres humanos durante a navegação do robô em três instantes de tempo diferentes (T_A, T_B e T_C).

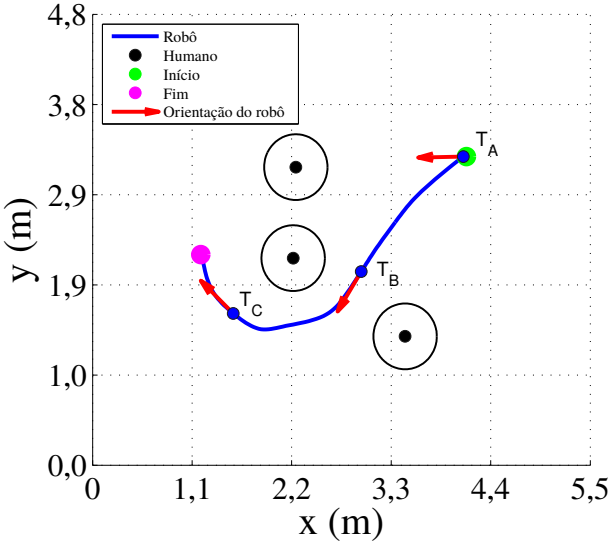
Figura 24 – Detecção durante a tarefa de desvio de seres humanos.



O Gráfico 8 mostra a trajetória descrita pelo robô e as posições dos seres humanos durante a navegação do robô. Usando a informação disponibilizada pelo serviço de detecção de seres humanos, o robô pode navegar sem tocar os indivíduos presentes no ambiente. Os instantes $\mathrm{T_A}$, $\mathrm{T_B}$ e $\mathrm{T_C}$, mostrados na Figura 24, são destacados ao longo da trajetória. A estratégia de planejamento de caminhos empregada procura pelo caminho com menor probabilidade de colisão.

Assim como apresentado na Seção 3.3.2.1, de forma a validar a efetividade do serviço proposto em atender aos requisitos da aplicação, uma análise da acurácia do detector nos planos da imagem e do chão é realizada, respectivamente, por meio das Tabelas 12 e 13. É

Gráfico 8 – Trajetória descrita pelo robô e as posições dos seres humanos durante a navegação do robô. O círculo ao redor do ser humano possui diâmetro igual a 0,5 m e define a área ocupada pelo corpo do ser humano no plano do chão.



possível observar um comportamento dos resultados semelhante ao observado no experimento da Seção 3.3.2.1. Em geral, o método proposto possui uma maior PR global, ao custo de uma maior MR. No entanto, essa última observação é atenuada pelo emprego de uma rede de câmeras e do esquema de memorização das últimas 10 posições tridimensionais do ser humano. É válido ressaltar que esse procedimento de memorização será avaliado na Seção 3.3.2.3.

Note que o baixo valor de FN, apresentado na Tabela 13, corrobora com o baixo valor de MR* (em negrito) apresentado na Tabela 12. É possível perceber que os detectores ACF e LDCF apresentam, novamente, maiores valores globais de FPPI. Embora esses detectores genéricos possuam valores absolutos de FPPI relativamente pequenos em algumas câmeras, eles não seriam capazes de empregar a rede de câmeras em toda a sua extensão, o que é crucial para configurações mais dinâmicas, como aquelas apresentadas nas Seções 3.3.2.1 e 3.3.2.3.

Por outro lado, a solução proposta nesta tese de doutorado é mais flexível, tendo em vista que não é dependente de uma câmera em específico, servindo, portanto, a uma ampla gama de aplicações. Note também que, o erro de localização para esse experimento é apenas 12% do diâmetro médio da área do corpo humano projetada no plano do chão (0,5 m). Esse erro de localização pode ser considerado pequeno, tendo em vista que o processo de detecção é aplicado a uma imagem de baixa resolução (515x291 píxeis). Finalmente, o serviço de detecção proposto foi capaz de cooperar com os demais elementos da infraestrutura, de forma a executar a tarefa com sucesso.

De forma adicional aos experimentos apresentados, nesta tese, relativos à tarefa de desvio de seres humanos, um conjunto de vídeos acerca do presente experimento é disponibilizado em (ALMONFREY, 2018).

Tabela 12 – Análise da detecção de seres humanos no plano da imagem para a segunda aplicação, considerando o serviço proposto e os detectores ACF e LDCF.

Dispositivos	Serviço proposto				ACF			LDCF		
	MR (%)	MR* (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)
Câmera 1	0	0	0	100	0	5,19	36,7	0	2,0	60,0
Câmera 2	2,5	0	$4,0 \cdot 10^{-2}$	98,5	0	7,36	29,0	0	3,3	47,3
Câmera 3	3,5	0	$4,0 \cdot 10^{-2}$	98,5	0	0,86	77,7	0	$3,2 \cdot 10^{-1}$	90,2
Câmera 4	37,4	0	$2,6 \cdot 10^{-1}$	82,8	0	1,93	51,0	0	$8,0 \cdot 10^{-2}$	96,1
Todas as Câmeras	8,4	**0**	$8,7 \cdot 10^{-2}$	**96,7**	0	3,83	41,8	0	1,4	65,7

*As BB de imagens de outras câmeras são projetadas na imagem da câmera analisada. Isso mostra a vantagem do emprego de uma rede de câmeras.

Tabela 13 – Análise, no plano do chão, do processo de detecção ao longo da segunda aplicação.

TP	FP	FN	NDTS	NGTS	LOCERR (m)
405	0	0	405	405	0,06

NDTS - Representa o número total de detecções;
NGTS - Esse é o número de anotações estimadas, no plano do chão, empregando as BB anotadas no plano da imagem;
LOCERR - Representa o erro médio de localização entre uma detecção positiva e sua anotação correspondente.

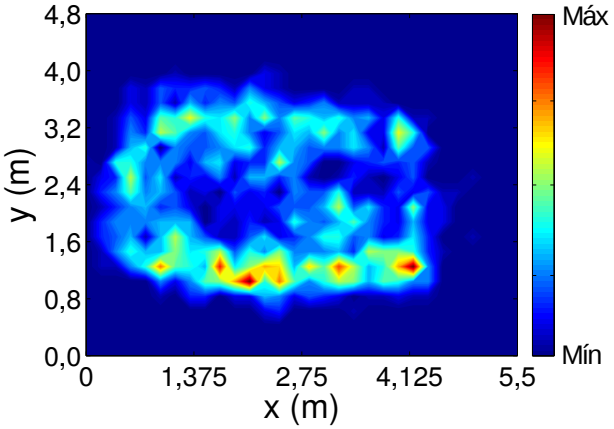
3.3.2.3 Mapa de ocupação cumulativo do ambiente

Neste experimento, um mapa de ocupação cumulativo do ambiente é construído. O objetivo é mostrar os locais mais visitados no ambiente ao longo do tempo. O Gráfico 9 mostra o mapa de ocupação cumulativo do ambiente. Esse mapa é construído por meio da contagem do número de seres humanos, que passaram sobre uma determinada posição no espaço, em um dado intervalo de tempo. É válido ressaltar que a identidade do indivíduo não foi considerada. O espaço inteligente mede 4,8 m x 7,3 m. No entanto, devido ao posicionamento do sistema de câmeras, parte do corpo dos indivíduos presentes no ambiente fica posicionado fora da imagem ou é distorcido pela maior parte das câmeras nos últimos 1,8 m do eixo horizontal (x). Como nenhum ser humano é detectado nessa região, apenas uma área de 4,8 m x 5,5 m é exibida. Essa região é considerada como a área de operação do espaço inteligente.

No experimento desta seção, solicitou-se aos indivíduos executar, principalmente, uma trajetória elíptica e isso pode ser confirmado por meio do Gráfico 9. Conforme mencionado anteriormente, em (SURIE; PARTONIA; LINDGREN, 2013), em que a solução é mais estruturada devido a limitação do sensor Kinect$^{\mathrm{TM}}$, os autores mencionam que um maior campo de visão poderia impactar o processo de detecção negativamente. Isso se deve ao fato do sistema desenvolvido estar atrelado às características do sensor utilizado. Como o sistema proposto nesta tese é menos estruturado, um maior campo de visão irá beneficiar a detecção, devido à maior área de operação.

Conforme já realizado em outros experimentos, as Tabelas 14 e 15 apresentam a análise de acurácia nos planos da imagem e do chão, respectivamente. Novamente, o comportamento observado nos resultados é similar aos dos demais experimentos, ao se comparar o serviço de detecção desenvolvido com os detectores ACF e LDCF. A diferença neste experimento é que mais pessoas estão presentes no ambiente (no máximo quatro ao mesmo tempo), aumentando a MR de todos os detectores, devido a maior taxa de oclusão. De fato, oclusão é ainda um

Gráfico 9 – Mapa de ocupação cumulativo do ambiente. As cores presentes na figura representam o grau de ocupação do plano do chão ao longo do tempo. Nesse sentido, as siglas Máx e Mín representam os valores relativos às ocupações máxima e mínima, respectivamente.



problema em aberto na literatura. É importante mencionar que, mesmo os melhores detectores genéricos apresentam uma redução no desempenho quando se considera um cenário com oclusão, conforme pode ser confirmado no repositório apresentado em (CALTECH, 2009).

Tabela 14 – Análise da detecção de seres humanos no plano da imagem para a terceira aplicação, considerando o serviço proposto e os detectores ACF e LDCF.

Dispositivos	Serviço proposto				ACF			LDCF		
	MR (%)	MR* (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)	MR (%)	FPPI	PR (%)
Camera 1	$38,3$	$26,2$	$5,5 \cdot 10^{-3}$	$93,5$	$4,5$	$4,95$	$20,0$	$5,4$	$0,53$	$69,7$
Camera 2	$48,5$	$26,3$	$8,4 \cdot 10^{-2}$	$87,1$	$7,9$	$5,10$	$16,8$	$9,9$	$1,58$	$39,0$
Camera 3	$60,2$	$22,3$	$5,1 \cdot 10^{-2}$	$86,5$	$13,5$	$0,67$	$51,5$	$17,4$	$0,20$	$77,6$
Camera 4	$84,4$	$29,8$	$1,8 \cdot 10^{-2}$	$85,6$	28.1	0.22	$69,7$	$36,5$	$7,0 \cdot 10^{-2}$	$86,5$
Todas as Câmeras	$54,0$	$\mathbf{26,0}$	$5,2 \cdot 10^{-2}$	$\mathbf{89,6}$	$11,6$	$2,73$	$24,26$	$14,7$	$0,59$	$58,7$

*As BB de imagens de outras câmeras são projetadas na imagem da câmera analisada. Isso mostra a vantagem do emprego de uma rede de câmeras.

Tabela 15 – Análise, no plano do chão, do processo de detecção ao longo da terceira aplicação.

TP	FP	FN	NDTS	NGTS	LOCERR (m)	FP*	FN*
1411	115	595	1526	2006	$0,18$	198	159

NDTS - Representa o número total de detecções;
NGTS - Esse é o número de anotações estimadas, no plano do chão, empregando as BB anotadas no plano da imagem;
LOCERR - Representa o erro médio de localização entre uma detecção positiva e sua anotação correspondente;
*Métricas calculadas após o relaxamento do limiar de distância euclidiana e o cômputo da média do vetor de memória das últimas 10 posições registradas para os seres humanos.

A partir da Tabela 15, é também possível notar que a MR no plano do chão $(\frac{\text{FN}}{\text{NGTS}})$ é aproximadamente 29%, o que está próximo à MR* global no plano da imagem, quando se considera informações advindas de todas as câmeras (Tabela 14). A precisão do serviço de detecção é maior que a dos demais detectores usados para comparação, como pode ser confirmado também por meio da Tabela 14.

De maneira a realizar uma análise complementar, é possível demonstrar que o alto número de FN, no plano do chão, está de certa forma relacionado a variações mínimas nas detecções, devido

ao cenário mais desafiador deste experimento. Somente nesta avaliação, o limiar de distância euclidiana será relaxado de 0,5 m para 0,8 m, tanto para o processo de agrupamento ilustrado na Figura 21, quanto para o processo de cômputo das métricas no plano do chão. Além disso, nesse caso, as métricas são calculadas após o emprego do vetor de memorização das 10 últimas posições do ser humano, de forma a avaliar a efetividade desse procedimento.

Como pode ser visto a partir da Tabela 15, quando se considera esse cenário mais relaxado, o número de FN* reduz consideravelmente ao custo de um mínimo aumento do número de FP*. Como pode ser visto, com uma consideração razoável acerca do cenário de avaliação, vários falsos negativos são evitados. Isso mostra que muitos erros são, de fato, cometidos por uma pequena margem, o que representa um mínimo impacto nas aplicações. Dessa forma, a maior MR no plano da imagem é, de fato, reduzida no plano do chão quando empregando uma rede de câmeras e o vetor de memorização.

No que diz respeito ao erro de localização, a partir da Tabela 15, pode-se notar um leve aumento no erro em relação ao apresentado nos outros experimentos. Isso é causado, principalmente, devido ao cenário mais dinâmico deste experimento, em que existe mais movimento e oclusão. Esse fato aumenta a instabilidade no processo de detecção em cada imagem e, consequentemente, no procedimento de localização no plano do chão. No entanto, uma vez que a detecção é aplicada a uma imagem de baixa resolução (515x291 píxeis), um erro menor que 40%, do diâmetro médio da área projetada pelo corpo humano no plano do chão, pode ser considerado razoável.

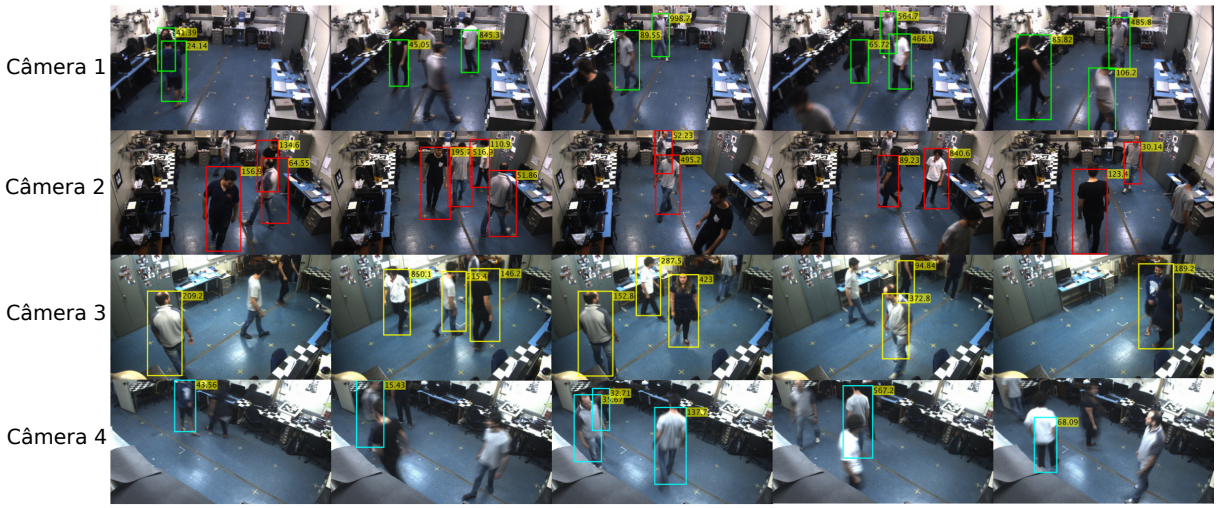
Finalmente, a Figura 25 apresenta alguns resultados qualitativos, incluindo alguns casos de sucesso e falha no processo de detecção para situações com oclusão. Vale ressaltar que, para uma maior diversidade de situações, as imagens das câmeras não são necessariamente correspondentes nesse caso. De forma adicional aos experimentos apresentados, nesta tese, relativos à criação do mapa de ocupação cumulativo do ambiente, um conjunto de vídeos acerca do presente experimento é disponibilizado em (ALMONFREY, 2018).

3.3.2.4 Erro de localização do serviço de detecção de seres humanos

Como uma avaliação final, calculou-se o erro de localização do serviço de detecção proposto para posições distribuídas em um grid uniforme, ao longo da área de operação do espaço inteligente (Gráfico 10). Para isso, um indivíduo foi posicionado em locais pré-estabelecidos, enquanto imagens eram capturadas. Essas posições foram medidas diretamente no plano do chão, de forma a se estabelecer as anotações para o cálculo das métricas empregadas. A ideia é usar esse cenário mais estruturado, de maneira a se realizar uma melhor análise sobre a influência do processo de detecção e também dos dados de calibração, no erro de localização.

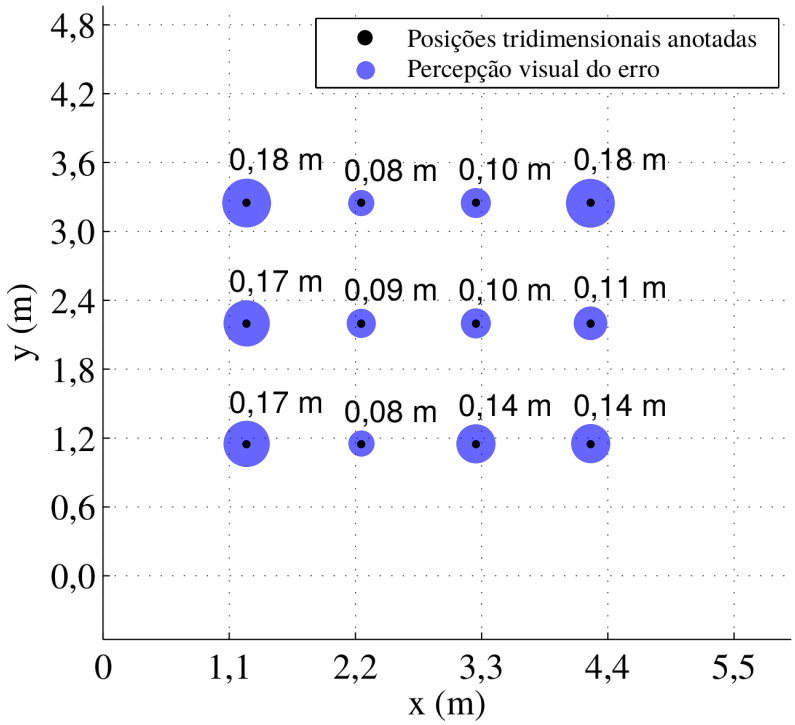
Como pode ser visto por meio do Gráfico 10, esse erro não é distribuído uniformemente ao longo do ambiente, e isso se deve, provavelmente, a redundância não uniforme inserida pelas áreas de sobreposição das câmeras do sistema utilizado. A redundância na detecção é também importante para uma mais precisa localização no plano do chão. Nas bordas da área de operação,

Figura 25 – Resultados qualitativos de detecção para o terceiro experimento. Podem ser notados casos de sucesso e falha no processo de detecção, em situações de oclusão.



onde existe uma menor redundância na detecção, o erro é maior. No entanto, o erro obtido é novamente menor que 40% do diâmetro médio da área do corpo de um ser humano projetada no plano do chão.

Gráfico 10 – Erro de localização ao longo da área de operação do espaço inteligente.



3.4 Conclusões deste capítulo

Neste capítulo, apresentou-se um serviço de detecção adequado para diferentes aplicações, em um espaço inteligente baseado em uma rede de câmeras. Uma revisão bibliográfica foi

realizada de forma a destacar as principais diferenças entre a solução proposta neste trabalho e aquelas existentes na área. Enquanto a maior parte dos trabalhos da literatura desenvolvem detectores de seres humanos como uma aplicação, a solução apresentada nesta tese é proposta como um serviço. Esse, por sua vez, é desenvolvido para interagir com a arquitetura de um espaço inteligente, de forma a prover as informações necessárias para as aplicações que o empregam.

Além disso, por ser implementado utilizando conceitos como computação em nuvem e arquitetura orientada a serviços (SOA), o serviço torna-se flexível, sendo capaz de atender a diferentes aplicações. A solução proposta provê, além de informações adequadas no que tange ao processo de detecção, características como paralelismo, escalabilidade e confiabilidade, existentes no âmbito do espaço inteligente no qual é desenvolvida.

Por fim, um sistema de câmeras é empregado para eliminar alguns problemas apresentados por detectores genéricos de seres humanos, que os fazem ser desconsiderados em situações práticas ou tarefas de tempo real. No que diz respeito ao tempo e a qualidade de detecção, o serviço se mostrou, por meio de experimentos, adequado para interagir com outros serviços e também com a infraestrutura do espaço inteligente, no sentido de executar as tarefas propostas.

4 Conclusões finais e trabalhos futuros

Nesta tese de doutorado, abordou-se o problema de detecção de seres humanos, no contexto de aplicações para um espaço inteligente baseado em uma rede de câmeras. Um procedimento de detecção foi proposto em uma arquitetura orientada a serviços (SOA), que emprega conceitos de computação em nuvem e possui características como escalabilidade, paralelismo e confiabilidade. A principal premissa utilizada foi a construção de uma solução flexível tanto em termos de implantação como usabilidade. Os detectores empregados são eficientes em termos computacionais e não dependentes de *hardware* específico, como unidades de processamento gráfico (GPU). Dessa forma, o serviço proposto pode ser instanciado em qualquer nó da infraestrutura física e ser empregado por diferentes aplicações de tempo real. Como contribuição adicional, propõe-se uma alteração no processo de extração de características, do fluxograma de operação de um detector de seres humanos, por meio do emprego da análise de componentes independentes (ICA).

Inicialmente, no Capítulo 2, o problema de detecção de seres humanos foi abordado no contexto da detecção de pedestres, devido à maturidade dessa área na comunidade científica. Um esquema de extração de características orientado aos dados, baseado na análise de componentes independentes (ICA), foi proposto. Técnicas de extração de características, orientadas aos dados, são formas mais consistentes de se obter descritores do que aquelas manualmente projetadas (*handcrafted*). Uma comparação no âmbito da família de soluções conhecida como canais de características filtradas (FCF) foi realizada. Essa família é conhecida pela eficiência do processo de detecção, possuindo soluções adequadas, para aplicações de tempo real, empregando apenas unidades centrais de processamento (CPU). Além disso, as soluções dessa família possuem uma boa relação custo computacional *versus* qualidade de detecção.

Uma comparação final com outros métodos existentes na literatura também foi conduzida para a base de dados Caltech. Dado que o objetivo era avaliar os pontos fortes e fracos do processo de extração de características, toda a análise foi realizada considerando modelos de classificadores pré-existentes na literatura. O classificador em questão é formado por um conjunto de árvores de decisão treinadas via AdaBoost. O processo de detecção proposto se mostrou competitivo com soluções da família analisada. Para um conjunto específico de dados anotados com maior precisão na base de dados Caltech, a metodologia de extração de características apresentou resultados próximos até mesmo de soluções baseadas em aprendizado profundo.

Em seguida, no Capítulo 3, um serviço de detecção foi proposto no âmbito de um espaço inteligente baseado em uma rede de câmeras. De forma a se ter um serviço com acoplamento mínimo à infraestrutura física, detectores independentes de GPU foram empregados. Os detectores escolhidos foram o ACF e ICCF, analisados no Capítulo 2. Além disso, o serviço proposto foi projetado para operar em uma arquitetura baseada em serviços, que emprega conceitos de computação em nuvem. Por meio da virtualização do detector em um contêiner Docker, a utilização de

um barramento de comunicação e um controlador de replicação (Kubernets), características como escalabilidade, confiabilidade e paralelismo são garantidas. É importante mencionar também que a arquitetura modular, do fluxograma do processo de detecção desenvolvido, possibilitou a utilização do paralelismo intrínseco ao paradigma empregado pelo espaço inteligente. Em termos de qualidade de detecção, alguns problemas, enfrentados por detectores genéricos de seres humanos, foram resolvidos ao se estruturar minimamente o espaço físico, por meio do uso de uma rede de câmeras. Foram utilizados conceitos de subtração de fundo e homografia, junto ao detector que emprega um modelo baseado em aparências, para implementar um eficiente procedimento em cascata para eliminar falsos positivos.

A validação do serviço proposto foi realizada por meio da análise de funcionamento de três aplicações provas de conceito (PdC), sendo que duas delas envolvem interação homem-máquina. Análises de requisitos como qualidade e tempo de detecção foram conduzidas e comparações com dois detectores genéricos de seres humanos foram realizadas. Isso permitiu realizar um paralelo entre soluções para aplicações de tempo real e aquelas que são apenas analisadas em bases de dados públicas, por meio de experimentos *off-line*. O serviço de detecção proposto se mostrou adequado para interagir com os demais serviços da arquitetura, de forma a atender as demandas das aplicações. Em termos de implementação e usabilidade, não se tem conhecimento, ao alcance da revisão bibliográfica realizada, de um serviço de detecção tal qual o desenvolvido nesta tese.

Como trabalhos futuros podem ser destacados:

- Integrar técnicas baseadas em aprendizado profundo (*deep learning*) ao serviço de detecção. Nesse caso, de forma a manter como característica a flexibilidade na alocação dos serviços, na infraestrutura física, deve-se desenvolver formas de evitar o emprego de GPU. A reimplementação de algum detector genérico existente na literatura, utilizando o paralelismo a nível de nós de processamento, pode ser explorada nesse caso;

- Análise dos efeitos dos ruídos de transmissão e compressão de informação, além do ruído relativo à variação da iluminação, no processo de detecção em tempo real e distribuído em uma rede;

- Troca da etapa de subtração de fundo por um procedimento de segmentação semântica, de forma a remover a restrição de fixação do sistema de câmeras;

- Desenvolvimento de um serviço *on-line* de recalibração, utilizando a informação de detecção dos seres humanos, para o caso de movimento do sistema de câmeras;

- Especificação e análise das restrições dos modelos de detecção de seres humanos, por meio do projeto de uma arquitetura distribuída utilizando uma abordagem *top-down*.

Referências

ADDUCI, M.; AMPLIANITIS, K.; REULKE, R. A quality evaluation of single and multiple camera calibration approaches for an indoor multi camera tracking system. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, p. 9–15, 2014. Citado na página 61.

AHMEDALI, T.; CLARK, J. J. Collaborative multi-camera surveillance with automated person detection. In: *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*. [S.l.: s.n.], 2006. p. 39–39. Citado na página 62.

AKKALADEVI, S. C.; HEINDL, C. Action recognition for human robot interaction in industrial applications. In: *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*. [S.l.: s.n.], 2015. p. 94–99. Citado na página 24.

ALAHI, A.; RAMANATHAN, V.; FEI-FEI, L. Chapter 6 - tracking millions of humans in crowded spaces. In: MURINO, V. et al. (Ed.). *Group and Crowd Behavior for Computer Vision*. Academic Press, 2017. p. 115 – 135. ISBN 978-0-12-809276-7. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128092767000072>. Citado na página 23.

ALBAWENDI, S. et al. Overview of behavioural understanding system with filtered vision sensor. In: *2015 International Conference on Interactive Technologies and Games*. [S.l.: s.n.], 2015. p. 90–95. Citado 3 vezes nas páginas 24, 61 e 75.

ALMONFREY, D. *Repositório com Vídeos dos Experimentos do Doutorado*. 2018. <https://bitbucket.org/Monfa/exps_doutorado/>. [Acesso em: 21 de maio de 2018.]. Citado 3 vezes nas páginas 90, 91 e 94.

ALMONFREY, D. et al. A flexible human detection service suitable for intelligent spaces based on a multi-camera network. *International Journal of Distributed Sensor Networks*, v. 14, n. 3, p. 1550147718763550, 2018. Disponível em: <https://doi.org/10.1177/1550147718763550>. Citado na página 28.

ALMONFREY, D. et al. *Human Detection Service*. 2018. <https://bitbucket.org/Monfa/humandetectionservice>. [Acesso em: 25 de abril de 2018.]. Citado na página 84.

ALMONFREY, D. et al. Modelo estatístico para filtragem de exemplos negativos na detecção de pedestres. In: . [S.l.: s.n.], 2016. <http://www.swge.inf.br/proceedings/paper/?P=CBA2016-0595>. Citado 2 vezes nas páginas 28 e 31.

ALMONFREY, D. et al. Neural cells insights on pedestrian detection. In: *XXI Congresso Brasileiro de Automática*. [S.l.: s.n.], 2016. <http://www.swge.inf.br/proceedings/paper/?P=CBA2016-0590>. Citado na página 28.

ANGELOVA, A. et al. Real-time pedestrian detection with deep network cascades. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2015. p. 32.1–32.12. ISBN 1-901725-53-7. Citado na página 35.

APPEL, R. et al. Quickly boosting decision trees: Pruning underachieving features early. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. JMLR.org, 2013. (ICML'13), p. III–594–III–602. Disponível em: <http://dl.acm.org/citation.cfm?id=3042817.3043003>. Citado 2 vezes nas páginas 43 e 44.

ATIS. *Telecom Glossary*. 2016. <http://www.atis.org/glossary/>. [Acesso em: 15 de março de 2018]. Citado na página 27.

BENENSON, R. et al. Seeking the strongest rigid detector. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2013. p. 3666–3673. ISSN 1063-6919. Citado na página 44.

BENENSON, R. et al. Ten years of pedestrian detection, what have we learned? In: AGAPITO, L.; BRONSTEIN, M. M.; ROTHER, C. (Ed.). *Computer Vision - ECCV 2014 Workshops*. Cham: Springer International Publishing, 2015. p. 613–627. ISBN 978-3-319-16181-5. Citado 6 vezes nas páginas 23, 25, 31, 32, 34 e 35.

BOUTSIS, I.; KALOGERAKI, V.; GUNO, D. Reliable crowdsourced event detection in smartcities. In: *2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in partnership with Global City Teams Challenge (GCTC) (SCOPE - GCTC)*. [S.l.: s.n.], 2016. p. 1–6. Citado na página 24.

BRAZIL, G.; YIN, X.; LIU, X. Illuminating pedestrians via simultaneous detection and segmentation. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2017. p. 4960–4969. Citado 2 vezes nas páginas 35 e 57.

BROGGI, A. et al. Shape-based pedestrian detection. In: *Proceedings of the IEEE Intelligent Vehicles Symposium 2000 (Cat. No.00TH8511)*. [S.l.: s.n.], 2000. p. 215–220. Citado na página 31.

BRSCIC, D. Social robots in smart public environments. In: *2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*. [S.l.: s.n.], 2014. p. 651–653. Citado na página 25.

BURNS, B. et al. Borg, omega, and kubernetes. *Queue*, ACM, v. 14, p. 10:70–10:93, 2016. Citado 2 vezes nas páginas 69 e 71.

CAI, Z. et al. A unified multi-scale deep convolutional neural network for fast object detection. In: LEIBE, B. et al. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 354–370. ISBN 978-3-319-46493-0. Citado 2 vezes nas páginas 59 e 72.

CAI, Z.; SABERIAN, M.; VASCONCELOS, N. Learning complexity-aware cascades for deep pedestrian detection. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 3361–3369. Citado 2 vezes nas páginas 31 e 34.

CALTECH. *Caltech Pedestrian Detection Benchmark*. 2009. <http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/>. [Acesso em: 15 de março de 2018.]. Citado 2 vezes nas páginas 45 e 93.

CANEDO-RODRIGUEZ, A. et al. Self-organized multi-camera network for a fast and easy deployment of ubiquitous robots in unknown environments. *Sensors*, MDPI AG, v. 13, n. 1, p. 426–454, Dec 2012. ISSN 1424-8220. Disponível em: <http://dx.doi.org/10.3390/s130100426>. Citado na página 62.

CAO, J.; PANG, Y.; LI, X. Pedestrian detection inspired by appearance constancy and shape symmetry. *IEEE Transactions on Image Processing*, v. 25, n. 12, p. 5538–5551, Dec 2016. ISSN 1057-7149. Citado 4 vezes nas páginas 25, 32, 34 e 57.

CAO, J.; PANG, Y.; LI, X. Learning multilayer channel features for pedestrian detection. *IEEE Transactions on Image Processing*, v. 26, n. 7, p. 3210–3220, July 2017. ISSN 1057-7149. Citado 2 vezes nas páginas 32 e 34.

CHEN, H. et al. Intelligent agents meet the semantic web in smart spaces. *IEEE Internet Computing*, v. 8, p. 69–79, 2004. Citado na página 24.

CHEN, S.-L.; CHANG, S.-K.; CHEN, Y.-Y. Development of a multisensor embedded intelligent home environment monitoring system based on digital signal processor and wi-fi. *International Journal of Distributed Sensor Networks*, v. 11, p. 171365, 2015. Citado na página 25.

CHRUNGOO, A.; MANIMARAN, S. S.; RAVINDRAN, B. Activity recognition for natural human robot interaction. In: BEETZ, M.; JOHNSTON, B.; WILLIAMS, M.-A. (Ed.). *Social Robotics*. Cham: Springer International Publishing, 2014. p. 84–94. ISBN 978-3-319-11973-1. Citado na página 24.

COEN, M. H. Design principles for intelligent environments. In: *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*. [S.l.: s.n.], 1998. p. 547–554. Citado na página 24.

COOK, D. J. et al. Detection of social interaction in smart spaces. *Cybernetics and systems*, v. 41, n. 2, p. 90–104, 2010. Citado na página 25.

COOK, D. J. et al. Casas: A smart home in a box. *Computer*, v. 46, n. 7, p. 62–69, 2013. Citado na página 25.

DALAL, N. *Finding People in Images and Videos*. Tese (Theses) — Institut National Polytechnique de Grenoble - INPG, jul. 2006. Disponível em: <https://tel.archives-ouvertes.fr/tel-00390303>. Citado 2 vezes nas páginas 33 e 44.

DENG, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. In: *CVPR09*. [S.l.: s.n.], 2009. Citado na página 57.

DOLLÁR, P. *Piotr's Computer Vision Matlab Toolbox*. 2016. <http://pdollar.github.io/toolbox/index.html>. [Acesso em: 15 de março de 2018.]. Citado 5 vezes nas páginas 39, 44, 47, 52 e 55.

DOLLÁR, P. et al. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 36, n. 8, p. 1532–1545, Aug 2014. ISSN 0162-8828. Citado 6 vezes nas páginas 34, 36, 57, 72, 74 e 84.

DOLLÁR, P.; APPEL, R.; KIENZLE, W. Crosstalk cascades for frame-rate pedestrian detection. In: FITZGIBBON, A. et al. (Ed.). *Computer Vision – ECCV 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 645–659. ISBN 978-3-642-33709-3. Citado na página 44.

DOLLÁR, P.; BELONGIE, S.; PERONA, P. The fastest pedestrian detector in the west. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2010. p. 68.1–68.11. ISBN 1-901725-40-5. Doi:10.5244/C.24.68. Citado 2 vezes nas páginas 23 e 36.

DOLLÁR, P. et al. Integral channel features. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2009. p. 91.1–91.11. ISBN 1-901725-39-1. Doi:10.5244/C.23.91. Citado 2 vezes nas páginas 33 e 44.

DOLLÁR, P. et al. Pedestrian detection: A benchmark. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 304–311. ISSN 1063-6919. Citado 2 vezes nas páginas 45 e 46.

DOLLÁR, P. et al. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 34, n. 4, p. 743–761, April 2012. ISSN 0162-8828. Citado 2 vezes nas páginas 35 e 45.

FELZENSZWALB, P.; MCALLESTER, D.; RAMANAN, D. A discriminatively trained, multiscale, deformable part model. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2008. p. 1–8. ISSN 1063-6919. Citado na página 33.

FREJLICHOWSKI, D. et al. Smartmonitor - an intelligent security system for the protection of individuals and small properties with the possibility of home automation. *Sensors*, v. 14, p. 9922–9948, 2014. Citado na página 24.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, n. 1, p. 119 – 139, 1997. ISSN 0022-0000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002200009791504X>. Citado na página 33.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.*, The Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 04 2000. Disponível em: <https://doi.org/10.1214/aos/1016218223>. Citado 3 vezes nas páginas 42, 43 e 44.

GERONIMO, D. et al. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 7, p. 1239–1258, July 2010. ISSN 0162-8828. Citado 2 vezes nas páginas 23 e 35.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.]: IEEE Computer Society, 2014. p. 580–587. Citado 2 vezes nas páginas 25 e 59.

GLAS., D. F. et al. Human-robot interaction in public and smart spaces. *Intelligent Assistive Robots: Recent Advances in Assistive Robotics for Everyday Activities*, Springer International Publishing, v. 106, p. 235–273, 2015. Citado 2 vezes nas páginas 23 e 25.

GLAS, D. F. et al. The network robot system: Enabling social human-robot interaction in public spaces. *Journal of Human-Robot Interaction.*, Journal of Human-Robot Interaction Steering Committee, v. 1, n. 2, p. 5–32, 2013. Citado 4 vezes nas páginas 23, 25, 60 e 61.

GOMES, R. L. et al. How can emerging applications benefit from eaas in open programmable infrastructures? In: *IEEE Summer School on Smart Cities 2017: IEEE S3C2017*. [S.l.: s.n.], 2017. Citado na página 66.

HELAL, S. et al. The gator tech smart house: a programmable pervasive space. *Computer*, v. 38, p. 50–60, 2005. Citado na página 24.

HOSANG, J. et al. Taking a deeper look at pedestrians. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 4073–4082. ISSN 1063-6919. Citado na página 34.

HU, Q. et al. Pushing the limits of deep cnns for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1–1, 2017. ISSN 1051-8215. Citado 3 vezes nas páginas 23, 32 e 34.

HYVÄRINEN, A. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision - Code.* 2015. <http://www.naturalimagestatistics.net/>. [Acesso em: 15 de março de 2018.]. Citado na página 40.

HYVÄRINEN, A.; HURRI, J.; HOYER, P. O. *Natural Image Statistics: a probabilistic approach to early computational vision*. [S.l.: s.n.], 2009. Vol. 39. Citado na página 40.

LEE, C.; NORDSTEDT, D.; HELAL, S. Enabling smart spaces with osgi. *IEEE Pervasive Computing*, v. 2, n. 3, p. 89–94, July 2003. Citado 2 vezes nas páginas 23 e 25.

LEE, J.-E. et al. Human and robot localization using histogram of oriented gradients (hog) feature for an active information display in intelligent space. *Advanced Science Letters*, v. 5, p. 1–8, 2012. Citado 3 vezes nas páginas 25, 61 e 62.

LEE, J. H. et al. Robust pedestrian detection by combining visible and thermal infrared cameras. *Sensors*, v. 15, n. 5, p. 10580–10615, 2015. ISSN 1424-8220. Disponível em: <http://www.mdpi.com/1424-8220/15/5/10580>. Citado na página 35.

LEE, J.-H.; HASHIMOTO, H. Intelligent space: concept and contents. *Advanced Robotics*, Taylor & Francis, v. 16, n. 3, p. 265–280, 2002. Citado 2 vezes nas páginas 24 e 64.

LI, B.; YAO, Q.; WANG, K. A review on vision-based pedestrian detection in intelligent transportation systems. In: *Proceedings of 2012 9th IEEE International Conference on Networking, Sensing and Control*. [S.l.: s.n.], 2012. p. 393–398. Citado 2 vezes nas páginas 25 e 59.

LIU, X. et al. Detecting and counting people in surveillance applications. In: *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* [S.l.: s.n.], 2005. p. 306–311. Citado na página 23.

LIU, Y.-F.; GUO, J.-M.; CHANG, C.-H. Low resolution pedestrian detection using light robust features and hierarchical system. *Pattern Recognition*, v. 47, n. 4, p. 1616 – 1625, 2014. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320313004585>. Citado na página 31.

MAO, J. et al. What can help pedestrian detection? In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 6034–6043. ISSN 1063-6919. Citado na página 59.

MATSUHIRA, N. et al. Development of robotic transportation system - shopping support system collaborating with environmental cameras and mobile robots -. In: *ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*. [S.l.: s.n.], 2010. p. 1–6. Citado 2 vezes nas páginas 23 e 61.

MATSUYAMA, T.; UKITA, N. Real-time multitarget tracking by a cooperative distributed vision system. *Proceedings of the IEEE*, v. 90, n. 7, p. 1136–1150, Jul 2002. ISSN 0018-9219. Citado na página 63.

MEHMOOD, M. O. *Détection de personnes pour des systèmes de videosurveillance multi-caméra intelligents*. Tese (Doutorado) — Ecole Centrale de Lille, 2015. Thèse de doctorat dirigée par Chainais, Pierre et Achard, Catherine Automatique, génie informatique, traitement du signal et images Ecole centrale de Lille 2015. Disponível em: <http://www.theses.fr/2015ECLI0016>. Citado na página 63.

MERKEL, D. Docker: Lightweight linux containers for consistent development and deployment. *Linux J.*, Belltown Media, v. 2014, mar. 2014. Citado na página 66.

MORIOKA, K.; HASHIKAWA, F.; TAKIGAWA, T. Human identification based on walking detection with acceleration sensor and networked laser range sensors in intelligent space. *International Journal on Smart Sensing and Intelligent Systems*, v. 6, n. 5, p. 2040–2054, 2013. Citado na página 25.

MRAZOVAC, B. et al. System design for passive human detection using principal components of the signal strength space. In: *2012 IEEE 19th International Conference and Workshops on Engineering of Computer-Based Systems*. [S.l.: s.n.], 2012. p. 164–172. Citado na página 25.

MUJA, M. et al. Rein - a fast, robust, scalable recognition infrastructure. In: *2011 IEEE International Conference on Robotics and Automation*. [S.l.: s.n.], 2011. p. 2939–2946. ISSN 1050-4729. Citado na página 63.

NAM, W.; DOLLÁR, P.; HAN, J. H. Local decorrelation for improved pedestrian detection. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 424–432. Disponível em: <http://dl.acm.org/citation.cfm?id=2968826.2968874>. Citado 8 vezes nas páginas 37, 38, 48, 49, 50, 52, 55 e 84.

NGUYEN, D. T.; LI, W.; OGUNBONA, P. O. Human detection from images and videos: A survey. *Pattern Recognition*, v. 51, p. 148 – 175, 2016. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320315003179>. Citado na página 23.

OHN-BAR, E.; TRIVEDI, M. M. To boost or not to boost? on the limits of boosted trees for object detection. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. [S.l.: s.n.], 2016. p. 3350–3355. Citado 7 vezes nas páginas 31, 35, 36, 48, 55, 56 e 58.

OUYANG, W.; WANG, X. Single-pedestrian detection aided by multi-pedestrian detection. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2013. p. 3198–3205. ISSN 1063-6919. Citado na página 35.

PAISITKRIANGKRAI, S.; SHEN, C.; HENGEL, A. v. d. Pedestrian detection with spatially pooled features and structured ensemble learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 38, n. 6, p. 1243–1257, June 2016. ISSN 0162-8828. Citado na página 31.

PAISITKRIANGKRAI, S.; SHEN, C.; HENGEL, A. van den. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: FLEET, D. et al. (Ed.). *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, 2014. p. 546–561. ISBN 978-3-319-10593-2. Citado na página 31.

PATIL, S.; TALELE, K. Suspicious movement detection and tracking based on color histogram. In: *2015 International Conference on Communication, Information Computing Technology (ICCICT)*. [S.l.: s.n.], 2015. p. 1–6. Citado na página 23.

PEREIRA, F. G.; VASSALLO, R. F.; SALLES, E. O. T. Human–robot interaction and cooperation through people detection and gesture recognition. *Journal of Control, Automation and Electrical Systems*, v. 24, n. 3, p. 187–198, Jun 2013. ISSN 2195-3899. Disponível em: <https://doi.org/10.1007/s40313-013-0040-3>. Citado na página 24.

PICORETI, R. *Desenvolvimento de plataforma baseada em computação em nuvem para aproveitamento de recursos ociosos* . Vitória: [s.n.], 2017. 60 p. Citado na página 66.

PREMEBIDA, C. et al. Pedestrian detection combining rgb and dense lidar data. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. [S.l.: s.n.], 2014. p. 4112–4117. ISSN 2153-0858. Citado na página 35.

PYO, Y. et al. Service robot system with an informationally structured environment. *Robot. Auton. Syst.*, North-Holland Publishing Co., Amsterdam, The Netherlands, The Netherlands, v. 74, n. PA, p. 148–165, dez. 2015. ISSN 0921-8890. Disponível em: <https://doi.org/10.1016/j.robot.2015.07.010>. Citado 2 vezes nas páginas 23 e 63.

QUEIROZ, F. M. de. *Desenvolvimento da Infraestrutura de um Espaço Inteligente baseado em Visão Computacional e IoT*. Vitória: [s.n.], 2016. 80 p. Citado na página 66.

RAMPINELLI, M. et al. An intelligent space for mobile robot localization using a multi-camera system. *Sensors*, v. 14, n. 8, p. 15039–15064, 2014. Citado 2 vezes nas páginas 65 e 79.

REDMON, J.; FARHADI, A. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. Disponível em: <http://arxiv.org/abs/1612.08242>. Citado na página 59.

REN, S. et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, v. 39, p. 1137–1149, 2017. Citado na página 59.

RIBEIRO, D. et al. A real-time deep learning pedestrian detector for robot navigation. In: *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. [S.l.: s.n.], 2017. p. 165–171. Citado 2 vezes nas páginas 23 e 59.

ROSTANSKI, M.; GROCHLA, K.; SEMAN, A. Evaluation of highly available and fault-tolerant middleware clustered architectures using rabbitmq. In: *2014 Federated Conference on Computer Science and Information Systems*. [S.l.: s.n.], 2014. p. 879–884. Citado na página 68.

SARKAR, C. et al. Diat: A scalable distributed architecture for iot. *IEEE Internet of Things Journal*, v. 2, n. 3, p. 230–239, June 2015. ISSN 2327-4662. Citado na página 63.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: . [s.n.], 2014. abs/1409.1556. Disponível em: <http://arxiv.org/abs/1409.1556>. Citado na página 72.

ŞUCAN, I. A.; MOLL, M.; KAVRAKI, L. E. The Open Motion Planning Library. *IEEE Robotics & Automation Magazine*, v. 19, n. 4, p. 72–82, December 2012. <http://ompl.kavrakilab.org>. Citado na página 90.

SURIE, D.; PARTONIA, S.; LINDGREN, H. Human sensing using computer vision for personalized smart spaces. In: *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*. [S.l.: s.n.], 2013. p. 487–494. Citado 3 vezes nas páginas 25, 60 e 92.

TIAN, Y. et al. Deep learning strong parts for pedestrian detection. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 1904–1912. Citado na página 34.

VARGA, D.; SZIRÁNYI, T. Robust real-time pedestrian detection in surveillance videos. *Journal of Ambient Intelligence and Humanized Computing*, v. 8, n. 1, p. 79–85, Feb 2017. ISSN 1868-5145. Disponível em: <https://doi.org/10.1007/s12652-016-0369-0>. Citado na página 23.

VIOLA, P.; JONES, M. J.; SNOW, D. Detecting pedestrians using patterns of motion and appearance. p. 734–741 vol.2, Oct 2003. Citado na página 32.

WALK, S. et al. New features and insights for pedestrian detection. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 1030–1037. ISSN 1063-6919. Citado na página 35.

WANG, G. et al. A scalable distributed architecture for intelligent vision system. *IEEE Transactions on Industrial Informatics*, v. 8, n. 1, p. 91–99, Feb 2012. ISSN 1551-3203. Citado na página 62.

WANG, S. et al. PCN: Part and context information for pedestrian detection with cnns. In: *Proceedings of the British Machine Vision Conference*. [S.l.]: BMVA Press, 2017. Citado na página 57.

WANG, T. et al. Intelligent systems for industrial robotics: application in logistic field. *Industrial Robot: An International Journal*, v. 39, p. 251–259, 2012. Citado na página 23.

WEISER, M. The computer for the 21st century. *Scientific American*, v. 265, n. 3, p. 66–75, January 1991. Citado na página 24.

WEISER, M. The world is not a desktop. *Interactions*, v. 1, n. 1, p. 7–8, jan 1994. ISSN 1072-5520. Citado na página 24.

WENG, X. et al. Rotational rectification network: Enabling pedestrian detection for mobile vision. In: *IEEE Winter Conf. on Applications of Computer Vision*. [S.l.: s.n.], 2018. Citado na página 34.

WRIGHT, S.; STEVENTON, A. Intelligent spaces — the vision, the opportunities and the barriers. *BT Technology Journal*, v. 22, n. 3, p. 15–26, Jul 2004. ISSN 1573-1995. Citado 2 vezes nas páginas 24 e 64.

WU, B.; NEVATIA, R. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, v. 75, n. 2, p. 247–266, Nov 2007. ISSN 1573-1405. Disponível em: <https://doi.org/10.1007/s11263-006-0027-7>. Citado na página 33.

YANG, B. et al. Convolutional channel features. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2015. p. 82–90. Citado 3 vezes nas páginas 32, 34 e 50.

ZABULIS, X. et al. Multicamera human detection and tracking supporting natural interaction with large-scale displays. *Machine Vision and Applications*, v. 24, n. 2, p. 319–336, Feb 2013. Citado na página 25.

ZHANG, L. et al. Is faster r-cnn doing well for pedestrian detection? In: LEIBE, B. et al. (Ed.). *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016. p. 443–457. ISBN 978-3-319-46475-6. Citado 6 vezes nas páginas 23, 32, 34, 57, 59 e 72.

ZHANG, S. et al. How far are we from solving pedestrian detection? In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 1259–1267. Citado 4 vezes nas páginas 23, 35, 45 e 56.

ZHANG, S.; BENENSON, R.; SCHIELE, B. Filtered channel features for pedestrian detection. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015. p. 1751–1760. ISSN 1063-6919. Citado 6 vezes nas páginas 23, 31, 32, 33, 39 e 50.

ZHANG, S.; BENENSON, R.; SCHIELE, B. Citypersons: A diverse dataset for pedestrian detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 4457–4465. ISSN 1063-6919. Citado 2 vezes nas páginas 59 e 73.

ZHANG, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 22, n. 11, p. 1330–1334, nov. 2000. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/34.888718>. Citado na página 68.

ZHAO, X. et al. Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification. *Pattern Recognition*, v. 48, n. 6, p. 1947 – 1960, 2015. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S003132031400524X>. Citado na página 35.

ZHOU, Z. et al. Towards omnidirectional passive human detection. In: *2013 Proceedings IEEE INFOCOM*. [S.l.: s.n.], 2013. p. 3057–3065. Citado na página 25.

ZHU, M.; ROZELL, C. J. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLOS Computational Biology*, Public Library of Science, v. 9, n. 8, p. 1–15, 08 2013. Disponível em: <https://doi.org/10.1371/journal.pcbi.1003191>. Citado na página 40.

ZYLBERBERG, J.; DEWEESE, M. R. Sparse coding models can exhibit decreasing sparseness while learning sparse codes for natural images. *PLOS Computational Biology*, Public Library of Science, v. 9, n. 8, p. 1–10, 08 2013. Disponível em: <https://doi.org/10.1371/journal.pcbi.1003182>. Citado na página 40.