**UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS**

Angelica Tiemi Mizuno Nakamura

**Avaliação e Proposta de Sistemas de Câmeras Estéreo
para Detecção de Pedestres em Veículos Inteligentes**

São Carlos

2017

Angelica Tiemi Mizuno Nakamura

**Avaliação e Proposta de Sistemas de Câmeras Estéreo
para Detecção de Pedestres em Veículos Inteligentes**

Dissertação apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para obtenção do título de Mestre em Ciências, Programa de Engenharia Elétrica.

Área de concentração: Sistemas Dinâmicos

Orientador: Prof. Dr. Valdir Grassi Junior

São Carlos

2017

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

N163a Nakamura, Angelica Tiemi Mizuno
Avaliação e Proposta de Sistemas de Câmeras Estéreo
para Detecção de Pedestres em Veículos Inteligentes /
Angelica Tiemi Mizuno Nakamura; orientador Valdir
Grassi Junior. São Carlos, 2017.

Dissertação (Mestrado) - Programa de Pós-Graduação
em Engenharia Elétrica e Área de Concentração em
Sistemas Dinâmicos -- Escola de Engenharia de São
Carlos da Universidade de São Paulo, 2017.

1. Detecção de Pedestres. 2. Visão Computacional.
3. Visão Estéreo. 4. Veículos Inteligentes. I. Título.

FOLHA DE JULGAMENTO

Candidata: Engenheira **ANGELICA TIEMI MIZUNO NAKAMURA**.

Título da dissertação: "Avaliação e proposta de sistemas de câmeras estéreo para detecção de pedestres em veículos inteligentes".

Data da defesa: 06/12/2017.

Comissão Julgadora:

Resultado:

Prof. Dr. **Valdir Grassi Junior**
(Orientador)
(Escola de Engenharia de São Carlos/EESC)

Aprovada

Profa. Dra. **Raquel Frizera Vassallo**
(Universidade Federal do Espírito Santo/UFES)

Aprovada

Prof. Dr. **Fernando Santos Osório**
(Instituto de Ciências Matemáticas e de Computação/ICMC-USP)

Aprovada

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **Luís Fernando Costa Alberto**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luís Fernando Costa Alberto**

Dedico este trabalho à minha família. Em especial, aos meus pais, Aquemi Mizuno Nakamura e Hiroshi Nakamura, ao meu irmão, Diego Keidi Mizuno Nakamura, e aos meus avós, Yoshimi Nakamura e Shinkichi Nakamura (in memoriam).

AGRADECIMENTOS

Primeiramente, agradeço à minha família. Em especial, aos meus pais, Hiroshi Nakamura e Aquemi Mizuno Nakamura, e ao meu irmão, Diego Keidi Mizuno Nakamura, que nunca mediram esforços para que eu chegasse até aqui e sempre me apoiaram. Vocês são o meu maior tesouro.

Prof. Dr. Valdir Grassi Jr., muito obrigada pela orientação e por contribuir com o aprendizado e crescimento que tive durante essa jornada.

Agradeço a todos que colaboraram com o desenvolvimento deste projeto. Ao LRM, por toda ajuda, auxílio e esclarecimentos de dúvidas. E também, ao Ricardo, Kenny, Júnior e Luis, por participarem na coleta do conjunto de dados.

Ao meu namorado, Luiz Ricardo Takeshi Horita, por todo apoio que vem fornecendo desde a minha época de graduação.

Aos integrantes do LASI, pela amizade, companhia e momentos de descontração.

Finalmente, agradeço aos membros da banca, Prof. Dr. Fernando Santos Osório e Profa. Dra. Raquel Frizera Vassallo, que deram conselhos e críticas construtivas.

Muito obrigada a todos que fizeram parte dessa jornada!

RESUMO

NAKAMURA, A. T. M. **Avaliação e Proposta de Sistemas de Câmeras Estéreo para Detecção de Pedestres em Veículos Inteligentes**. 2017. 85p. Dissertação (Mestrado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Detecção de pedestres é uma importante área em visão computacional com o potencial de salvar vidas quando aplicada em veículos. Porém, essa aplicação exige detecções em tempo real, com alta acurácia e menor quantidade de falsos positivos possível. Durante os últimos anos, diversas ideias foram exploradas e os métodos mais recentes que utilizam arquiteturas profundas de redes neurais possibilitaram um grande avanço nesta área, melhorando significativamente o desempenho das detecções. Apesar desse progresso, a detecção de pedestres que estão distantes do veículo continua sendo um grande desafio devido às suas pequenas escalas na imagem, sendo necessária a avaliação da eficácia dos métodos atuais em evitar ou atenuar a gravidade dos acidentes de trânsito que envolvam pedestres. Dessa forma, como primeira proposta deste trabalho, foi realizado um estudo para avaliar a aplicabilidade dos métodos estado-da-arte para evitar colisões em cenários urbanos. Para isso, a velocidade e dinâmica do veículo, o tempo de reação e desempenho dos métodos de detecção foram considerados. Através do estudo, observou-se que em ambientes de tráfego rápido ainda não é possível utilizar métodos visuais de detecção de pedestres para assistir o motorista, pois nenhum deles é capaz de detectar pedestres que estão distantes do veículo e, ao mesmo tempo, operar em tempo real. Mas, ao considerar apenas pedestres em maiores escalas, os métodos tradicionais baseados em janelas deslizantes já conseguem atingir um bom desempenho e rápida execução. Dessa forma, com a finalidade de restringir a operação dos detectores apenas para pedestres em maiores escalas e assim, possibilitar a aplicação de métodos visuais em veículos, foi proposta uma configuração de câmeras que possibilitou obter imagens para um maior intervalo de distância à frente do veículo com pedestres em resolução quase duas vezes maior em comparação à uma câmera comercial. Resultados experimentais mostraram considerável melhora no desempenho das detecções, possibilitando superar a dificuldade causada pelas pequenas escalas dos pedestres nas imagens.

Palavras-chave: Detecção de Pedestres. Visão Computacional. Visão estéreo. Veículos Inteligentes.

ABSTRACT

NAKAMURA, A. T. M. **Stereo Cameras Systems Evaluation and Proposal for Pedestrian Detection on Intelligent Vehicles**. 2017. 85p. Dissertação (Mestrado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Pedestrian detection is an important area in computer vision with the potential to save lives when applied on vehicles. This application requires accurate detections and real-time operation, keeping the number of false positives as minimal as possible. Over the past few years, several ideas were explored, including approaches with deep network architectures, which have reached considerably better performances. However, detecting pedestrians far from the camera is still challenging due to their small sizes on images, making it necessary to evaluate the effectiveness of existing approaches on avoiding or reducing traffic accidents that involves pedestrians. Thus, as the first proposal of this work, a study was done to verify the state-of-the-art methods applicability for collision avoidance in urban scenarios. For this, the speed and dynamics of the vehicle, the reaction time and performance of the detection methods were considered. The results from this study show that it is still not possible to use a vision-based pedestrian detector for driver assistance on urban roads with fast moving traffic, since none of them is able to handle real-time pedestrian detection. However, for large-scale pedestrians on images, methods based on sliding window approach can already perform reliably well with fast inference time. Thus, in order to restrict the operation of detectors only for pedestrians in larger scales and enable the application of vision-based methods in vehicles, it was proposed a camera setup that provided to get images for a larger range of distances in front of the vehicle with pedestrians resolution almost twice as large compared to a commercial camera. Experimental results reveal a considerable enhancement on detection performance, overcoming the difficulty caused by the reduced scales that far pedestrians have on images.

Keywords: Pedestrian Detection. Computer Vision. Stereo Vision. Intelligent Vehicles.

LISTA DE FIGURAS

Figura 1 –	Projeção perspectiva do modelo de câmera <i>pinhole</i>	25
Figura 2 –	Modelo de câmera perspectiva.	26
Figura 3 –	Relação entre o sistema de coordenadas do mundo e da câmera.	27
Figura 4 –	Imagens obtidas por duas câmeras com diferentes pontos de vista.	29
Figura 5 –	Sistema estéreo composto por duas câmeras <i>pinholes</i> com os planos de imagens esquerda e direita coplanares.	30
Figura 6 –	Métodos de correspondência baseados em correlação (quadrados) e em características (círculos). Métodos baseado em correlação: regiões na imagem de tamanho fixo são comparadas. Métodos baseado em características: características relevantes são extraídas das imagens e comparadas.	31
Figura 7 –	Geometria epipolar	32
Figura 8 –	Retificação de par de imagens.	33
Figura 9 –	Etapas de execução de um sistema de detecção de pedestres baseado em janelas deslizantes.	34
Figura 10 –	Ilustração da abordagem tradicional para detectar pedestres em multi- escalas.	35
Figura 11 –	Ilustração do <i>stixel world</i> composto pelo chão plano e colunas verticais que representam os obstáculos.	36
Figura 12 –	Exemplos de imagens com os <i>ground truth</i> dos <i>datasets</i> INRIA, Caltech- USA e KITTI.	37
Figura 13 –	Ilustração de um classificador forte composto por árvores de decisão de nível 2 com três classificadores <i>stump</i> . Dada uma imagem de entrada, os canais de características (histograma de 6 orientações de gradientes, 1 magnitude de gradiente e 3 canais de cores LUV) são computados e então, é utilizado a técnica de imagens integrais para obter regiões aleatórias dos canais e alimentar as árvores de decisão de nível 2.	38
Figura 14 –	Comparação de desempenho dos detectores no INRIA <i>Pedestrian Data-</i> <i>set</i> , com imagens de tamanho 640×480 <i>pixels</i>	39
Figura 15 –	Esquerda: mapa de disparidade (quanto maior a disparidade, maior o nível de cinza). Direita: <i>v-disparity</i> correspondente ao mapa de disparidade.	40
Figura 16 –	Diagrama de operação do detector <i>VeryFast</i> com <i>stixels</i>	42

Figura 17 – Resultados dos algoritmos avaliados pelo <i>Caltech Pedestrian Detection Benchmark</i> no <i>dataset</i> da Caltech, com atualização realizada em Dezembro de 2016. (a) Desempenho absoluto realizado em todo o <i>dataset</i> . (b-c) Razão de aspecto típica e atípica. (d) Pedestres maiores que 80 <i>pixels</i> . (e) Pedestres com altura entre 30 a 80 <i>pixels</i> . (f) Pedestres com altura menor que 30 <i>pixels</i> . (g) Pedestre sem oclusão. (h) Pedestres com oclusão parcial (1 ~ 35% da área oclusa). (i) Pedestres com oclusão densa (35 a 80% da área oclusa).	47
Figura 18 – Distância entre o sistema de visão e a frente do veículo.	51
Figura 19 – Estimativas da distância total percorrida por um veículo em situações de frenagens bruscas, considerando o tempo de reação do detector <i>VeryFast</i> (50 <i>Hz</i>) em condições de pista seca e molhada.	54
Figura 20 – Formação de imagem em uma máquina fotográfica.	54
Figura 21 – Distância vs. altura do pedestre na imagem (<i>Bumblebee</i> XB3).	55
Figura 22 – Comparação da distância percorrida durante o tempo de reação de um ser humano e de um sistema de detecção de pedestres.	56
Figura 23 – Sistema de câmeras proposta instalada na plataforma CaRINA II. . .	59
Figura 24 – Configuração do sistema de câmeras proposto para detecção de pedestres (dimensões em metros).	60
Figura 25 – Diagrama de operação para cada câmera estéreo do sistema proposto. .	60
Figura 26 – Distância vs. altura do pedestre na imagem em relação a diferentes distâncias focais. (a) Distância focal de 4,5 <i>mm</i> . (b) Distância focal de 8,5 <i>mm</i>	61
Figura 27 – Efeito de corte causado pela diferença de tamanho entre a lente e o sensor. .	62
Figura 28 – Diferentes regiões de interesse utilizadas para cada sistema estéreo. . .	63
Figura 29 – Calibração das câmeras	63
Figura 30 – Resultado da calibração. (a-b) Imagens esquerda e direita retificadas. (c) Mapa de disparidade.	65
Figura 31 – Pedestres rotulados em <i>frames</i> equivalentes obtidos pelas câmeras com distâncias focais de 4,5 <i>mm</i> (esquerda) e 8,5 <i>mm</i> (direta).	66
Figura 32 – <i>Precision</i> e <i>Recall</i>	67
Figura 33 – Exemplos de detecção de pedestres que estão até 34 metros de distância das câmeras nas imagens esquerda de cada sistema estéreo (da esquerda para a direita, <i>Bumblebee</i> , 4,5 <i>mm</i> e 8,5 <i>mm</i>). <i>Bounding boxes</i> (retângulos) na cor azul representam os <i>ground truth</i> , enquanto que na cor verde indicam detecções VPs, e na cor vermelha os FPs.	70
Figura 34 – Exemplo do resultado da detecção em situação de pedestres a uma distância de aproximadamente 50 <i>m</i> à frente das câmeras. Esquerda: <i>Bumblebee</i> . Direita: sistema proposto.	72

Figura 35 – Situações onde o detector falhou em uma determinada sequência de <i>frames</i>	73
Figura 36 – Resultado das detecções de pedestres com <i>stixels</i>	73

LISTA DE TABELAS

Tabela 1 – Parâmetros e valores usados para o cálculo da distância de frenagem. .	53
Tabela 2 – Relação entre a distância mínima que os pedestres devem ser detectados e sua altura na imagem para que seja possível realizar a frenagem a tempo.	55
Tabela 3 – Resultados das detecções - pedestres até 34 metros distância das câmeras.	69
Tabela 4 – Resultados das detecções - pedestres em uma região de 34 <i>m</i> a 50 <i>m</i> à frente das câmeras.	71

Lista de Abreviaturas

ABS		Anti-lock Braking System
ACF		Aggregate Channel Features
Adaboost		Adaptive Boosting
ADAS		Advanced Driver Assistance System
AOV		Angle of View
BBOX		Bounding Box
BSD		Berkeley Software Distribution
Caltech		California Institute of Technology
CaRINA		Carro Robótico Inteligente para Navegação Autônoma
ChnFtrs		Channel Features
CNN		Convolutional Neural Networks
CPU		Central Processing Unit
EESC		Escola de Engenharia de São Carlos
ETH		Eidgenössische Technische Hochschule
FN		Falso Negativo
FP		Falso Positivo
FPDW		Fastest Pedestrian Detector in the West
FPS		Frames por Segundo
F-DNN+SS	..	Fused Deep Neural Network with Semantic Segmentation
GPU		Graphical Processing Unit
GT		Ground Truth
HOG		Histograma de Gradientes Orientados
ICF		Integral Channel Features
ICMC		Instituto de Ciências Matemáticas e de Computação

INRIA		Institut National de Recherche en Informatique et en Automatique
LASI		Laboratório de Sistemas Inteligentes
LRM		Laboratório de Robótica Móvel
MATLAB		MATrix LABoratory
MR		Miss Rate
MS-CNN		Multi-scale CNN
OpenCV		Open Source Computer Vision Library
PPS		Pedestrian Protection System
R-CNN		Regions with CNN Features
ROI		Região de Interesse
ROS		Robot Operating System
SA-FastRCNN		Scale-Aware Fast R-CNN
SAD		Sum of Absolute Differences
SVM		Support Vector Machine
USP		Universidade de São Paulo
VP		Verdadeiro Positivo

SUMÁRIO

1	INTRODUÇÃO	21
2	CONCEITOS FUNDAMENTAIS	25
2.1	Modelo de câmera perspectiva	25
2.1.1	Parâmetros extrínsecos	26
2.1.2	Parâmetros intrínsecos	27
2.2	Visão estéreo	28
2.2.1	Correspondência	30
2.2.2	Geometria epipolar	32
2.2.3	Retificação	32
2.3	Métodos visuais de detecção de pedestres	33
2.3.1	Arquitetura	33
2.3.2	Características	34
2.3.3	Pirâmide de imagem	35
2.3.4	Classificadores	36
2.3.5	Segmentação de <i>foreground</i>	36
2.3.6	Dataset	37
2.4	VeryFast	38
2.4.1	Exploração de informações de profundidade	40
2.5	Considerações finais	42
3	REVISÃO BIBLIOGRÁFICA	43
3.1	Trabalhos Relacionados	43
3.2	Desempenho dos detectores	46
3.3	Considerações finais	49
4	APLICABILIDADE DOS MÉTODOS	51
4.1	Metodologia	51
4.1.1	Distância de frenagem	52
4.1.2	Distância de aquisição de imagem	52
4.1.3	Distância de reação	52
4.1.4	Distância de atuação do sistema de freio	52
4.2	Estudo de caso	53
4.3	Considerações finais	56
5	PROPOSTA DE CONFIGURAÇÃO DE CÂMERAS ESTÉREO	59
5.1	Visão geral do sistema	59

5.1.1	Distância focal equivalente	61
5.1.2	Região de interesse	62
5.2	Calibração das câmeras	63
5.3	Dataset	65
5.4	Considerações finais	66
6	RESULTADOS	67
6.1	Metodologia de avaliação	67
6.2	Experimentos e avaliação	68
6.2.1	Parte 1	68
6.2.2	Parte 2	71
6.3	Considerações finais	74
7	CONCLUSÃO	75
7.1	Trabalhos futuros	76
	REFERÊNCIAS	77
	APÊNDICES	83
	APÊNDICE A – PUBLICAÇÃO	85

1 INTRODUÇÃO

Durante os últimos anos, os acidentes automotivos se tornaram uma importante causa de fatalidades, sendo o atropelamento o tipo de acidente mais comum. Segundo o relatório divulgado pela Companhia de Engenharia de Tráfego (2017), somente no município de São Paulo no ano de 2016, houve mais de 19 mil vítimas em acidentes de trânsito. E dentre os acidentes fatais, mais de 40,2% das vítimas eram pedestres.

Devido a essa grande quantidade de acidentes, tanto as empresas automobilísticas quanto a comunidade científica têm buscado o desenvolvimento de sistemas inteligentes com a finalidade de evitar ou atenuar a gravidade dos acidentes de trânsito.

Um desses sistemas é conhecimento como *Advanced Driver Assistance System* (ADAS), que auxilia o motorista na tomada de decisões, fornece alertas em situações de risco, e provê medidas de segurança entre veículos e a faixa da pista. Ainda neste seguimento de assistência ao motorista, o *Pedestrian Protection System* (PPS) é um sistema que foca na segurança dos pedestres. Seu objetivo é detectar a presença de pessoas paradas ou em movimento em uma área específica ao redor do veículo em movimento, a fim de alertar o condutor, executar ações de frenagem, ou acionar *airbags* externos se uma colisão é inevitável (GERÓNIMO; LÓPEZ, 2014). Além dos sistemas assistivos ao motorista, há também os veículos autônomos, que poderão reduzir os acidentes de trânsito provocados por fatores humanos. No entanto, em ambos sistemas, é crucial que o pedestre seja detectado antecipadamente, para que a possibilidade do veículo conseguir frear a tempo seja maior.

No ramo da pesquisa de detecção de pedestres, alguns autores propõe o uso de câmeras ou combinações delas com outros tipos de sensores (SZARVAS; SAKAI; OGATA, 2006; PREMEBIDA; LUDWIG; NUNES, 2009; GONZÁLEZ et al., 2016; ROSERO, 2017). No entanto, apesar da câmera ser sensível a condições climáticas e de iluminação do ambiente (e.g. chuva ou visão noturna), ela é o tipo de sensor que contém a maior resolução angular (GANDHI; TRIVEDI, 2007), além de possuir informações de cor, forma, textura, e até mesmo profundidade, que são relevantes para detecção de pedestres. Por este motivo, neste trabalho foram abordados apenas os métodos visuais.

Em visão computacional, a detecção de pedestres é um ramo específico da detecção de objetos que consiste em encontrar regiões na imagem que contenham pedestres. Devido ao seu potencial para salvar vidas, extensas pesquisas são realizadas desde o final da década de 90, e algumas indústrias automotivas já empregam sistemas de detecção de pedestres em seus carros de luxo (e.g. Volvo, Mercedes-Benz, Toyota e Ford). Porém, apesar da considerável melhora dos métodos nos últimos anos, ainda há uma lacuna

entre o desempenho humano e os métodos estado-da-arte de detecção de pedestres, não sendo possível obter um alto desempenho (*precision* e *recall*) e execução em tempo real simultaneamente (DOLLÁR et al., 2009; ZHANG et al., 2016). Além disso, ainda não se sabe a eficácia dos métodos existentes para evitar acidentes.

Dessa forma, o presente trabalho propõe uma metodologia para avaliar a aplicabilidade dos métodos estado-da-arte de detecção de pedestres no contexto real de cenários urbanos, considerando a dinâmica e velocidade do veículo durante a frenagem e as limitações atuais de *hardware* e *software* que o sistema possui. Através desta metodologia, observou-se que, para um veículo a 60 km/h conseguir frear a tempo de evitar uma colisão, os métodos devem detectar os pedestres que estão a pelo menos 32 m de distância à frente do veículo, o que significa detectar pedestres com pequenas escalas nas imagens. Porém, como mostrado no *benchmark* da Caltech, o principal desafio dos métodos estado-da-arte consiste em detectar pedestres em menores escalas, onde o melhor método avaliado deixa de detectar 77% dos pedestres contidos no *dataset*. Além dessa dificuldade, os recentes métodos que obtêm maior taxa de acerto são todos baseados em arquiteturas profundas de redes neurais, e consequentemente, possuem um alto custo computacional, o que é indesejável em ambientes urbanos e com veículos percorrendo em maior velocidade.

Entretanto, os métodos tradicionais de detecção baseados em janelas deslizantes já conseguem obter um bom desempenho em situações de pedestres com maiores resoluções, atingindo taxas de erro praticamente nulas e com rápida execução (DOLLÁR et al., 2009). Assim, se fosse possível restringir a operação dos detectores apenas para pedestres com maiores escalas nas imagens, até mesmo aqueles que estão distantes do veículo, não seria necessário o uso de arquiteturas profundas de redes neurais. Ao invés disso, um método rápido e robusto baseado em janelas deslizantes e classificador linear seria uma melhor alternativa.

Seguindo essa premissa, esta dissertação tem como segunda proposta uma configuração de câmeras estéreo que possibilita a aquisição de imagens com pedestres em maiores escalas e em um maior intervalo de distância à frente do veículo. Isto foi possível através da combinação de dois pares de câmeras estéreo com diferentes distâncias focais e *baselines*.

Para analisar a eficácia do sistema proposto e comparar o desempenho com uma câmera comercial, um novo conjunto de dados foi coletado utilizando a configuração de câmeras proposta em sincronização com a *Bumblebee XB3* da Point Grey. Esta coleta de dados foi realizada com o auxílio da plataforma CaRINA II, um veículo autônomo em desenvolvimento do Laboratório de Robótica Móvel (LRM) no Instituto de Ciências Matemáticas e de Computação (ICMC-USP). A avaliação de desempenho foi realizada utilizando o algoritmo de detecção *VeryFast* proposto em Benenson et al. (2012b), mostrando que através da configuração de câmeras proposta foi possível obter resultados promissores, e com isso, superar o problema que os métodos enfrentam em relação a detectar pedestres

que estão distantes do veículo devido sua pequena escala na imagem.

O restante desta dissertação está organizada da seguinte maneira. O Capítulo 2 fornece uma fundamentação teórica em visão computacional e em métodos visuais de detecção de pedestres. No Capítulo 3 são apresentados os trabalhos relacionados. A análise da aplicabilidade dos métodos estado-da-arte em detecção de pedestres é realizado no Capítulo 4, e a proposta de uma nova configuração de câmeras é apresentada no Capítulo 5. Por fim, os resultados são apresentados no Capítulo 6 e o trabalho é concluído no Capítulo 7.

2 CONCEITOS FUNDAMENTAIS

Com o objetivo de fornecer uma breve fundamentação teórica, este capítulo apresenta alguns conceitos e ferramentas importantes para este projeto. Nas Seções 2.1 e 2.2 são explicados o modelo de câmera perspectiva e visão estéreo, tomando como base os livros de Trucco e Verri (1998) e Szeliski (2010). Na Seção 2.3 são apresentados conceitos fundamentais referentes à detecção de pedestres baseado em visão. E na Seção 2.4 é apresentado o detector *VeryFast* proposto por Benenson et al. (2012b), que foi utilizado neste trabalho durante a avaliação de desempenho.

2.1 Modelo de câmera perspectiva

O modelo de câmera perspectiva é uma aproximação matemática baseada no modelo *pinhole*, que descreve a relação entre as coordenadas de um ponto 3D no mundo e a sua projeção no plano da imagem (π), não levando em consideração as distorções geométricas ou desfoque causados por lentes e aberturas de tamanho finito (Figura 1).

Este modelo consiste de um plano de imagem π' e do centro de projeção $\mathbf{O}$ (*pinhole*). O eixo cZ que cruza perpendicularmente o plano da imagem no ponto $\mathbf{o}'$ (centro da imagem) é denominado como eixo óptico. Conforme pode ser observado na Figura 2, a imagem de $\mathbf{P}$, $\mathbf{p}$, é o ponto onde a reta $\mathbf{OP}$ cruza o plano de imagem π' . Ao considerar o sistema de coordenadas 3D da câmera com origem em $\mathbf{O}$, o plano da imagem π' ortogonal ao eixo óptico, e definindo $\mathbf{P} = [{}^cX, {}^cY, {}^cZ]^T$ e $\mathbf{p} = [x, y, f]^T$, é possível obter as equações de

Figura 1: Projeção perspectiva do modelo de câmera *pinhole*.

Fonte: Elaborado pela autora.

Figura 2: Modelo de câmera perspectiva.

Fonte: Adaptado de Trucco e Verri (1998).

projeção perspectiva no sistema de coordenadas da câmera:

$$\begin{aligned} x &= f \frac{{}^c X}{{}^c Z} \\ y &= f \frac{{}^c Y}{{}^c Z} \end{aligned} \quad (2.1)$$

Para realizar a reconstrução 3D ou computar a posição de um objeto no espaço, é necessária uma equação que relaciona as coordenadas de um ponto 3D no mundo com os seus pontos correspondentes na imagem. Nesta equação é assumido o conhecimento dos parâmetros extrínsecos e intrínsecos da câmera.

2.1.1 Parâmetros extrínsecos

Os parâmetros extrínsecos são os parâmetros geométricos que definem a pose do sistema de coordenadas da câmera em relação ao sistema de coordenadas conhecido no mundo. Esta pose pode ser representada pelas seguintes variáveis:

- Vetor de translação 3D, ${}^w\mathbf{T}_c$, que descreve a posição da origem do sistema de coordenadas da câmera em relação ao sistema de coordenadas do mundo;

$${}^w\mathbf{T}_c = [T_x, T_y, T_z]^T \quad (2.2)$$

- Matriz ortogonal 3×3 de rotação, ${}^c_w R$, que descreve as rotações necessárias para tornar os eixos do sistema de coordenadas da câmera paralelos aos eixos no sistema de coordenadas do mundo.

$${}^c_w R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1^T \\ \mathbf{R}_2^T \\ \mathbf{R}_3^T \end{bmatrix} \quad (2.3)$$

Figura 3: Relação entre o sistema de coordenadas do mundo e da câmera.

Fonte: Elaborado pela autora.

Dessa forma, como pode ser observado na Figura 3, a relação das coordenadas do ponto $\mathbf{P}$ entre o sistema de coordenadas do mundo, ${}^w\mathbf{P}$, e da câmera, ${}^c\mathbf{P}$, é dada por:

$${}^c\mathbf{P} = {}^c_wR({}^w\mathbf{P} - {}^w\mathbf{T}_c), \quad (2.4)$$

ou,

$${}^c\mathbf{P} = \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & -\mathbf{R}_1^T \cdot {}^w\mathbf{T}_c \\ r_{21} & r_{22} & r_{23} & -\mathbf{R}_2^T \cdot {}^w\mathbf{T}_c \\ r_{31} & r_{32} & r_{33} & -\mathbf{R}_3^T \cdot {}^w\mathbf{T}_c \end{bmatrix}}_{M_{ext}} \cdot \begin{bmatrix} {}^wX \\ {}^wY \\ {}^wZ \\ 1 \end{bmatrix}$$

2.1.2 Parâmetros intrínsecos

Os parâmetros intrínsecos são o conjunto de parâmetros que relacionam as coordenadas de um ponto na imagem com as coordenadas 3D dos pontos representados no referencial da câmera. Para o modelo de câmera perspectiva, são utilizados quatro conjuntos de parâmetros intrínsecos:

- Distância focal f : a distância entre o plano de imagem, π , e a origem da câmera, $\mathbf{O}$;

- (o_x, o_y) : coordenadas em *pixel* da posição o' na imagem;
- (s_x, s_y) : tamanho vertical e horizontal do *pixel* em milímetros;
- k_1 : distorção radial causada pela lente.

Entretanto, quando regiões de interesse são definidas nas imagens (como é o caso da Seção 5.1.2), os *pixels* periféricos são descartados e a distorção radial causada pela lente pode ser desconsiderada (TRUCCO; VERRI, 1998). Dessa forma, a conversão das coordenadas métricas de um ponto projetado no plano da imagem em relação à câmera para coordenadas em *pixels* pode ser descrita pelas seguintes equações:

$$\begin{aligned} x &= -(x_{im} - o_x)s_x \\ y &= -(y_{im} - o_y)s_y \end{aligned} \quad (2.5)$$

Então, através das Equações 2.1 e 2.5, é possível obter:

$$\begin{bmatrix} x_1 \\ x_2 \\ {}^cZ \end{bmatrix} = \underbrace{\begin{bmatrix} -f/s_x & 0 & o_x \\ 0 & -f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix}}_{M_{int}} \cdot {}^c\mathbf{P} \quad (2.6)$$

Substituindo 2.4 em 2.6, obtêm-se a matriz de projeção perspectiva:

$$\begin{bmatrix} x_1 \\ x_2 \\ {}^cZ \end{bmatrix} = M_{int} M_{ext} \begin{bmatrix} {}^wX \\ {}^wY \\ {}^wZ \end{bmatrix}, \quad (2.7)$$

ou,

$$\begin{bmatrix} x_1 \\ x_2 \\ {}^cZ \end{bmatrix} = \begin{bmatrix} -f/s_x & 0 & o_x \\ 0 & -f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & -\mathbf{R}_1^T \cdot {}^w\mathbf{T}_c \\ r_{21} & r_{22} & r_{23} & -\mathbf{R}_2^T \cdot {}^w\mathbf{T}_c \\ r_{31} & r_{32} & r_{33} & -\mathbf{R}_3^T \cdot {}^w\mathbf{T}_c \end{bmatrix} \begin{bmatrix} {}^wX \\ {}^wY \\ {}^wZ \end{bmatrix}.$$

onde,

$$\begin{cases} x_1 = x_{im} {}^cZ \\ x_2 = y_{im} {}^cZ \end{cases}$$

2.2 Visão estéreo

A visão estéreo consiste na capacidade de inferir a estrutura tridimensional a partir de imagens capturadas por duas ou mais câmeras que estão observando o mesmo cenário. No entanto, há dois problemas em sistemas de visão estéreo que devem ser resolvidos, a

Figura 4: Imagens obtidas por duas câmeras com diferentes pontos de vista.

Fonte: Elaborado pela autora.

correspondência e a reconstrução. Considerando um sistema de visão estéreo composto de uma câmera ao lado da outra, o primeiro problema consiste em determinar qual *pixel* da imagem esquerda corresponde ao *pixel* da imagem direita, e ao mensurar a distância entre esses dois pontos, é obtida a disparidade. Porém, como pode ser observado na Figura 4, um caso típico em visão estéreo é que as duas câmeras obtêm imagens de um mesmo cenário de diferentes pontos de vista, dessa forma, algumas partes da cena serão visíveis apenas por uma câmera. O segundo problema a ser resolvido é a reconstrução de uma representação 3-D da cena. Após solucionar o problema da correspondência, é possível calcular as disparidades de todos os pontos da imagem, formando o mapa de disparidade. Então, conhecendo a geometria do sistema estéreo, o mapa de disparidade pode ser convertido em um mapa 3-D.

Antes de continuar com os problemas de correspondência e reconstrução, será feita uma breve introdução sobre o modelo de sistema estéreo. Para isso, suponha um sistema estéreo composto por duas câmeras idênticas do tipo *pinhole*, onde seus eixos de projeções são paralelos, como mostrado na Figura 5. Os planos de imagem esquerda e direita (π_l e π_r) são coplanares entre si, sendo $\mathbf{O}_l$ e $\mathbf{O}_r$ seus centros de projeção. A distância entre os dois centros é denominado *baseline* (b) do sistema estéreo. Considerando um ponto $\mathbf{P}$ no mundo e suas respectivas projeções, $\mathbf{p}_l$ e $\mathbf{p}_r$, x_l e x_r serão as coordenadas de $\mathbf{p}_l$ e $\mathbf{p}_r$ em relação à posição dos pontos $\mathbf{o}_l$ e $\mathbf{o}_r$ na imagem, f é a distância focal e Z é a distância

Figura 5: Sistema estéreo composto por duas câmeras *pinholes* com os planos de imagens esquerda e direita coplanares.

Fonte: Elaborado pela autora.

entre o ponto $\mathbf{P}$ e a *baseline*. Através da triangulação entre $p_l P p_r$ e $O_l P O_r$ é possível obter:

$$Z = f \frac{b}{d}, \quad (2.8)$$

onde $d = x_r - x_l$ é a disparidade. De acordo com a Equação 2.8, é possível notar que a profundidade é inversamente proporcional à disparidade. E, ao considerar uma mesma distância (Z), o tamanho da *baseline* (b) atua como um fator de ampliação na mensuração de d para obter Z . Em outras palavras, a distância estimada é mais precisa se as duas câmeras forem configuradas mais distantes uma da outra, o que significa ter uma maior *baseline*. Porém, pelo fato de *baselines* mais longas resultarem em uma maior disparidade, as buscas por pontos correspondentes tornam-se mais desafiadores, podendo causar falsas correspondências (OKUTOMI; KANADE, 1993). Dessa forma, deve haver um balanceamento entre precisão e acurácia no momento de definir o tamanho da *baseline*.

2.2.1 Correspondência

Conforme é mostrado na Figura 4, quando duas câmeras observam um elemento no ambiente, ele é projetado para ambas imagens obtidas. Porém, normalmente o cenário é complexo, e encontrar pontos correspondentes nas imagens se torna uma tarefa difícil. Dessa forma, o problema de correspondência consiste em, dado um elemento na imagem

Figura 6: Métodos de correspondência baseados em correlação (quadrados) e em características (círculos). Métodos baseado em correlação: regiões na imagem de tamanho fixo são comparadas. Métodos baseado em características: características relevantes são extraídas das imagens e comparadas.

Fonte: Elaborada pela autora.

esquerda, encontrar o elemento correspondente na imagem direita. Trucco e Verri (1998) classificam os algoritmos de correspondência em duas classes, os métodos baseados em características e os baseados em correlação, conhecido também como métodos locais (Figura 6).

Os métodos baseados em características limitam a procura por correspondências a um conjunto de características. Neste método são utilizados descritores de características (normalmente de bordas, quinas ou *blobs*) para posteriormente medir a distância entre as características das imagens. Dessa forma, os elementos correspondentes consistem no par de características com maior similaridade, e que está associado à distância mínima. Alguns descritores de características utilizados são o SIFT, SURF, FAST e Harris (FRAUNDORFER; SCARAMUZZA, 2012).

Nos métodos baseados em correlação, uma janela de tamanho fixo é utilizada para encontrar elementos correspondentes nas duas imagens. Para isso, um critério de similaridade é usado para medir a correspondência entre as duas janelas nas imagens. Essa similaridade normalmente é medida utilizando os algoritmos *Sum of Absolute Differences* (SAD), *Sum of Squared Differences* (SSD) ou *Normalized Cross Correlation* (NCC) (HIRSCHMÜLLER; SCHARSTEIN, 2007). No pacote *stereo_image_proc* do ROS é utilizado o algoritmo de correspondência SAD que computa a diferença de todos os *pixels* da janela definida e em seguida soma o módulo de todos os resultados dessas diferenças. Assim, a janela resultante que tiver o menor valor será a que tem maior similaridade.

2.2.2 Geometria epipolar

Os algoritmos de correspondência permitem que pontos correspondentes nas imagens sejam encontrados. Porém, explorar parâmetros de calibração para realizar a retificação possibilita que a busca por regiões correspondentes seja realizada apenas ao longo da linha epipolar. Isto é possível devido a geometria da visão estéreo (geometria epipolar) que relaciona geometricamente os pontos 3D e suas projeções nas imagens (TRUCCO; VERRI, 1998; SZELISKI, 2010). Como exemplo, a Figura 7 ilustra duas câmeras observando o mesmo ponto P , e como o *pixel* desse ponto P na imagem esquerda, p_l , é projetada na outra imagem a partir de um segmento de linha epipolar. Esse segmento é delimitado em uma extremidade pelo ponto P na imagem, e na outra extremidade, pela projeção do centro da câmera O_l na segunda câmera, conhecida como epipolo (e_r). Dessa forma, as linhas epipolares são a interseção dos dois planos de imagens com o plano epipolar que passa pelo ponto P e pelos dois centros de câmeras, O_l e O_r (SZELISKI, 2010).

Figura 7: Geometria epipolar

Fonte: Elaborado pela autora.

2.2.3 Retificação

Se a linha epipolar não é conhecida, a busca por pontos correspondentes deve ser realizada em toda a imagem, o que é custoso computacionalmente. Porém, conhecendo os parâmetros de calibração do sistema estéreo é possível realizar a retificação das imagens obtidas pelas duas câmeras, simplificando o problema de correspondência para uma busca unidimensional.

Tendo um par de imagens estéreo, a retificação consiste em realizar transformações sobre os sistemas de coordenadas de cada câmera de forma que as linhas epipolares de ambas as imagens se tornem colineares e paralelas a um dos eixos das imagens, normalmente ao eixo horizontal (Figura 8). Para isso, Trucco e Verri (1998) definem o algoritmo de

Figura 8: Retificação de par de imagens.

Fonte: Elaborado pela autora.

retificação em quatro passos. Primeiro, rotaciona-se a câmera esquerda de forma a levar o epipolo ao infinito ao longo do eixo horizontal. Em seguida, aplica-se a mesma rotação para a câmera direita com o intuito de recuperar a geometria original. Feito isso, a câmera direita é rotacionada por R para tornar as câmeras paralelas entre si. Finalmente, realiza-se o ajuste de escala em ambas as imagens para garantir a colinearidade das linhas epipolares.

Quando as imagens são corretamente retificadas, a procura por pontos correspondentes é reduzido para uma única linha na imagem. Então, para cada *pixel* na imagem esquerda é pesquisada a melhor correspondência em sua linha epipolar na imagem direita.

2.3 Métodos visuais de detecção de pedestres

Embora várias diferentes ideias tenham sido exploradas no ramo da detecção de pedestres, e a qualidade da detecção ter sido melhorada, ainda não se sabe a eficácia e aplicabilidade dos métodos propostos até o momento para evitar ou atenuar a gravidade dos acidentes que envolvam pedestres. Bem como, se a busca por melhores características, métodos de classificação, exploração de informações adicionais (profundidade e movimento), dentre outros, é a direção correta a ser seguida para que em um futuro próximo seja possível solucionar o problema de detecção de pedestres para veículos.

Com a finalidade de proporcionar ao leitor uma familiarização com os termos e métodos que serão mencionados adiante, nesta seção serão apresentados alguns conceitos fundamentais referentes a detecção de pedestres baseado em visão.

2.3.1 Arquitetura

Em métodos visuais de detecção de pedestres, considera-se uma imagem como o dado de entrada. Para detectores baseados em janelas deslizantes, uma janela espacial de tamanho fixo é percorrida por toda a imagem realizando a extração de características

para cada posição da janela, e então, essas características são enviadas ao classificador para realizar a decisão entre a existência de pedestre ou não. Caso haja um pedestre em uma determinada posição da imagem, é retornada uma caixa delimitadora (*bounding box*). Dessa forma, ao final, a imagem terá um conjunto de caixas delimitadoras, representando as posições dos pedestres nela contida. Este sistema de detecção pode ser resumido nas etapas de execução mostradas na Figura 9.

Figura 9: Etapas de execução de um sistema de detecção de pedestres baseado em janelas deslizantes.

Fonte: Elaborado pela autora.

Considerando a aplicação em veículos inteligentes, as imagens podem sofrer forte influência da variação repentina de iluminação (e.g. sombras, túneis, etc.) e do movimento do cenário causado pelo deslocamento do veículo. Apesar de haver câmeras que ajustam automaticamente seus parâmetros (e.g. tempo de exposição e sensibilidade do sensor) de acordo com a condição de iluminação do cenário, ainda não há uma câmera que realize este ajuste rápido o suficiente. Por isso, costuma-se adotar a etapa de pré-processamento, na qual é realizado um tratamento na imagem para reduzir problemas relacionados a baixo contraste e efeitos indesejados causados pela iluminação do cenário, além de ressaltar as características importantes para as próximas etapas da detecção. Porém, como essa etapa fornece um ganho muito pequeno enquanto é requerido um maior custo computacional, alguns autores passaram a pular esta etapa (DALAL; TRIGGS, 2005).

Quanto ao processo de treinamento e classificação, geralmente é feito por um algoritmo de aprendizado de máquina supervisionado. Dessa forma, a partir de um conjunto de dados rotulados treina-se o sistema para reconhecer padrões das características extraídas de imagens, de forma a reconhecer as que melhor descrevem os pedestres. Após o treinamento, o sistema deverá ser capaz de reconhecer regiões das imagens que correspondam a classe de pedestres e não-pedestres.

2.3.2 Características

Características em uma imagem são partes locais, significativas e detectáveis, que diferem dos *pixels* vizinhos em termos de cor, intensidade ou textura. Como a aparência de uma pessoa varia bastante e não possui um formato padrão, em métodos de detecções de pedestres se torna necessário determinar uma característica robusta que seja informativa, invariante a ruídos, escala, translação, iluminação, e de rápida computação, de forma a

distinguir claramente uma pessoa de outros objetos em um cenário (DALAL; TRIGGS, 2005; DOLLÁR et al., 2007).

Durante os últimos anos, vários tipos de características foram exploradas para detecção de pedestres, tais como informações de cor, borda, textura, modelo local, características de covariância, dentre outros (DALAL; TRIGGS, 2005; LEIBE; SEEMANN; SCHIELE, 2005; WANG; HAN; YAN, 2009; WALK et al., 2010; MARIN et al., 2013; PAISITKRIANGKRAI; SHEN; HENGEL, 2013; COSTEA; NEDEVSKI, 2014). Atualmente, os métodos de detecção de pedestres com abordagem em janelas deslizantes que obtêm melhor desempenho são extensões do detector de canais de características integrais (*ChnFtrs*), que utilizam como características a combinação de canais de orientações de gradientes, magnitude de gradiente e canais de cor LUV (DOLLÁR et al., 2009; DOLLÁR; BELONGIE; PERONA, 2010; BENENSON et al., 2012b; DOLLÁR et al., 2014; ZHANG; BENENSON; SCHIELE, 2015; BENENSON et al., 2015).

2.3.3 Pirâmide de imagem

A arquitetura mais comum em detecções de pedestres é com abordagem em janelas deslizantes. Em geral, essa janela tem um tamanho fixo, o que causaria problemas ao tentar detectar pessoas com diferentes escalas em uma imagem. Desse modo, é realizada uma variação na resolução da imagem para obter escalas menores ou maiores da mesma e possibilitar a detecção de pedestres em multi-escalas. Para isso, um único modelo é treinado para uma escala regular, e então, a imagem de entrada é redimensionada N vezes (aproximadamente 50 vezes) (Figura 10). Apesar dessa abordagem se mostrar eficaz, ela possui problemas devido a necessidade de determinar a escala ideal para treinar o modelo e o elevado custo computacional, dado que as características devem ser computadas para cada imagem redimensionada (BENENSON et al., 2012b).

Figura 10: Ilustração da abordagem tradicional para detectar pedestres em multi-escalas.

Fonte: Elaborado pela autora.

2.3.4 Classificadores

Em detecção de pedestres, os classificadores lineares, tais como SVM linear ou Floresta de Decisão são os mais frequentemente utilizados devido a sua garantia teórica, extensibilidade e bom desempenho. No entanto, após a rede neural convolutiva (CNN) proposta por Krizhevsky, Sutskever e Hinton (2012) obter notáveis resultados na competição de classificação do ImageNet em 2012, iniciou-se uma nova onda de redes neurais para visão computacional (PEPIKBO et al., 2015).

2.3.5 Segmentação de *foreground*

A segmentação de *foreground* é uma técnica utilizada para delimitar regiões de interesse (ROIs) na imagem de entrada, de forma a evitar o máximo possível as regiões de *background* e reduzir o custo computacional.

Em Badino, Franke e Pfeiffer (2009) foi introduzido o *stixel world*, que é uma representação dos obstáculos por um conjunto retangular de colunas, denominadas *stixels* (Figura 11). A ideia básica é que o chão é localmente plano e todos os objetos podem ser descritos como um obstáculo que cresce verticalmente sobre o chão. Esta técnica é relevante não apenas pelo fato de reduzir o número de candidatos, mas também, por evitar regiões como o chão, céu ou áreas da imagem onde a profundidade é muito irregular (e.g., vegetação). Porém, um ponto muito importante desta etapa é garantir que todos os pedestres contidos na imagem estejam dentro da ROI delimitada, caso contrário, os módulos seguintes não serão capazes de corrigir este erro, e consequentemente, haverá um aumento na quantidade de falsos negativos (GERÓNIMO et al., 2010).

Figura 11: Ilustração do *stixel world* composto pelo chão plano e colunas verticais que representam os obstáculos.

Fonte: Benenson, Timofte e Van Gool (2011).

2.3.6 Dataset

Segundo Dollár et al. (2012), os conjuntos de dados (*dataset*) existentes podem ser agrupados em dois tipos, o *dataset* de pessoas e o de pedestres. O primeiro consiste de pessoas em diversas poses sob um vasto domínio, enquanto o segundo contém pessoas em pé (paradas ou andando), normalmente de pontos de vista mais restrito e muitas vezes contendo informações de movimento. Dentre os diversos *datasets* existentes, os mais frequentemente utilizados em detecção de pedestres são o KITTI (GEIGER; LENZ; URTASUN, 2012), Caltech-USA (DOLLÁR et al., 2009) e INRIA (DALAL; TRIGGS, 2005).

O INRIA é um dos *datasets* mais antigos e contém relativamente poucas imagens, capturadas através de fotografias de férias em vários cenários (cidade, praia, montanha, etc.). Suas imagens são de alta resolução e as anotações de pedestres normalmente são

Figura 12: Exemplos de imagens com os *ground truth* dos *datasets* INRIA, Caltech-USA e KITTI.

(a) INRIA

(b) Caltech-USA

(c) KITTI

Fonte: Elaborado pela autora a partir das imagens dos respectivos *datasets*.

incompletas. Já os *datasets* da KITTI e Caltech são atualmente os conjuntos de dados mais completos e desafiadores em detecção de pedestres, onde ambos contêm imagens capturadas através de um veículo em movimento e proveem informações estéreo (Figura 12).

2.4 VeryFast

O algoritmo de detecção de pedestres usado neste trabalho foi o *VeryFast* (BENENSON et al., 2012b), pois dentre os métodos estado-da-arte baseados em janelas deslizantes, ele é o mais rápido. Uma análise mais detalhada da razão pela qual o *VeryFast* foi escolhido para este trabalho será explicado nos Capítulos 3 e 4.

Nos métodos tradicionais de detecção de pedestres um único modelo de escala fixa é treinado, e a imagem de entrada é redimensionada N vezes para possibilitar a detecção em multi-escalas. Embora esses métodos se mostrem eficazes, eles apresentam dois problemas. Primeiro, é necessário encontrar uma escala intermediária para o modelo, de forma que este consiga detectar pedestres tanto em pequenas escalas (baixa resolução), quanto em maiores escalas (alta resolução). O segundo problema está relacionado ao tempo de processamento, que é elevado devido ao redimensionamento da imagem de entrada em N escalas e à computação de características para cada uma delas.

Figura 13: Ilustração de um classificador forte composto por árvores de decisão de nível 2 com três classificadores *stump*. Dada uma imagem de entrada, os canais de características (histograma de 6 orientações de gradientes, 1 magnitude de gradiente e 3 canais de cores LUV) são computados e então, é utilizada a técnica de imagens integrais para obter regiões aleatórias dos canais e alimentar as árvores de decisão de nível 2.

Fonte: Benenson et al. (2012b).

Com o intuito de solucionar esses dois problemas, Benenson et al. (2012b) propuseram um detector baseado em orientações de gradientes que não necessita redimensionar a imagem de entrada. Para isso, os autores basearam-se na ideia inversa do trabalho de Dollár, Belongie e Perona (2010), onde foi proposto um método híbrido que constrói amostras de pirâmide de imagem com intervalos de uma oitava, e dentro de cada oitava, é realizada a computação aproximada de características em escalas adjacentes para construir a pirâmide de classificadores. Em outras palavras, Benenson et al. (2012b) utilizaram apenas uma escala de imagem de entrada e cinco modelos treinados, e então, foi utilizado uma aproximação para transformar esses modelos treinados em N modelos, eliminando a necessidade de criar pirâmide de imagens.

Basicamente, um classificador forte é construído através de um conjunto de árvores de decisão, onde cada árvore contém três classificadores *stump*. Cada classificador *stump* é definido por um índice de canal, um retângulo sobre tal canal e um *threshold* de decisão (Figura 13). Assim, quando um *stump* é redimensionado por um fator s , o índice de canal se mantém constante, o retângulo é redimensionado por s , e o *threshold* é atualizado para $\tau' = \tau \cdot r(s)$. Onde $r(s)$ é o fator de variação da resposta de uma característica devido ao seu redimensionamento por um fator s . Com isso, são treinados N/K modelos (~ 5), sendo um para cada oitava.

Na etapa de testes, cada modelo treinado é convertido em K modelos para diferentes escalas, totalizando N modelos (~ 50). Posteriormente, os canais de características são computados na imagem de entrada, e então, as respostas para cada escala são computadas utilizando os N classificadores. De acordo com a avaliação realizada posteriormente em Dollár et al. (2014) utilizando o *dataset* do INRIA, o detector *VeryFast* mostrou-se mais interessante em comparação aos outros métodos avaliados, sendo o detector mais rápido e com o melhor desempenho, obtendo uma taxa de erro de 16% a 50 *Hz* (Figura 14). Apesar

Figura 14: Comparação de desempenho dos detectores no INRIA *Pedestrian Dataset*, com imagens de tamanho 640×480 *pixels*.

Fonte: Dollár et al. (2014).

do INRIA *Pedestrian Dataset* ser menos desafiador que outros *datasets* de pedestres, Dollár et al. (2012) mostrou que o desempenho relativo entre detectores independe do *dataset* utilizado.

2.4.1 Exploração de informações de profundidade

O uso de informações de profundidade em detecção de pedestres pode ser um forte aliado para evitar regiões com baixa probabilidade de haver obstáculos, tais como chão e céu. Porém, computar mapas de profundidade é uma tarefa muito demorada para operação em tempo real. Dessa forma, baseado no trabalho de Badino, Franke e Pfeiffer (2009), Benenson, Timofte e Van Gool (2011) propuseram um novo método onde os obstáculos que estão acima do chão são moldados diretamente a partir das imagens estéreo. Isto é, para cada coluna na imagem, são estimados o *pixel* inferior, superior e a distância até o obstáculo sem a necessidade de um mapa intermediário de profundidade. Para acelerar o tempo de execução, Benenson et al. (2012b) utilizaram o modelo de *stixel world* em combinação com o detector *VeryFast*. Então, com a posição central de um provável pedestre na imagem e sua possível escala, o detector irá procurar apenas em alguns *pixels* acima e abaixo da coluna e em algumas escalas. Isto possibilitou realizar detecções em imagens de tamanho 640×480 *pixels* a uma taxa de 135 *Hz*.

Segundo Benenson et al. (2012a), apesar dessa estimativa direta do *stixels* proporcionar uma significativa aceleração, ela acaba se tornando o fator limitante de velocidade quando é realizada a combinação com o detector em uma GPU. Por essa razão, os autores propuseram uma nova parametrização para estimativa direta do *stixels world* que permitiu

Figura 15: Esquerda: mapa de disparidade (quanto maior a disparidade, maior o nível de cinza). Direita: *v-disparity* correspondente ao mapa de disparidade.

Fonte: Labayrade, Aubert e Tarel (2002).

alcançar aproximadamente 165 Hz , sem comprometer a qualidade da detecção.

O modelo do *stixel world* proposto por Benenson et al. (2012a) pode ser decomposto em três conjuntos de parâmetros: *ground plane*, distância até os objetos e alturas dos objetos. A estimativa do *ground plane* consiste em encontrar o primeiro obstáculo relevante visível no sentido positivo de profundidade, ou seja, a procura deve começar na região inferior da imagem na direção vertical, até que uma célula ocupada seja encontrada. Dessa forma, o espaço encontrado antes dessa célula é considerado como espaço livre (*ground plane*). Este cálculo é realizado utilizando o método de *v-disparity* proposto no trabalho de Labayrade, Aubert e Tarel (2002), mas, sem a necessidade de computar o mapa de profundidade.

O mapa de disparidade é computado a partir da geometria epipolar, então por meio do acúmulo de *pixels* com a mesma disparidade em cada linha, é gerado o mapa *v-disparity*. Devido à distância da câmera, normalmente a parte inferior do mapa de profundidade tem um valor maior de disparidade, e a superior um valor menor. Então, o *ground plane* pode ser representado como uma linha diagonal no mapa de disparidade *v*, e o obstáculo é representado por uma linha vertical (Figura 15). A fim de acelerar essa computação, Benenson et al. (2012a) propuseram realizar esse acúmulo apenas uma vez a cada N linhas, proporcionando uma significativa aceleração sem a perda de qualidade em casos de imagens em maior resolução.

Benenson et al. (2012a) assumem que a altura dos objetos são conhecidos e dependentes da classe, então, após estimar o *ground plane*, a distância até os objetos é estimada diretamente a partir do domínio $u - v$, ao invés de realizar uma programação dinâmica no domínio do *u-disparity* (KUBOTA; NAKANO; OKAMOTO, 2007; BENENSON; TIMOFTE; Van Gool, 2011). Os códigos do algoritmo *VeryFast* e do *stixels* em domínio $u - v$ utilizados neste trabalho estão disponibilizados online por Benenson et al. (2012b) e Benenson et al. (2012a).

O algoritmo *VeryFast* combinado com *stixels* utiliza como entrada um par de imagens fornecida pela câmera estéreo para estimar a localização dos obstáculos e realizar as detecções (BENENSON et al., 2012b). O diagrama de seu funcionamento pode ser observado na Figura 16, onde, os parâmetros de calibração são utilizados para retificar as imagens obtidas pela câmera estéreo e estimar a localização dos obstáculos. Com isso, o detector opera apenas dentro da região onde foi considerada a existência de obstáculos, e caso haja a presença de pedestres na imagem, as posições dos pedestres identificados são retornadas.

Figura 16: Diagrama de operação do detector *VeryFast* com *stixels*.

Fonte: Elaborado pela autora.

2.5 Considerações finais

Neste capítulo foram fornecidos alguns conceitos fundamentais de visão computacional que foram utilizados para o desenvolvimento deste trabalho e apresentados os principais termos e processos dos métodos visuais de detecção de pedestres.

O estudo de modelo de câmera e visão estéreo foram importantes para obter a altura do pedestre na imagem, estimar sua distância do veículo e determinar a configuração de câmeras estéreo proposta neste trabalho. Para avaliar a contribuição do sistema proposto em relação à melhora no desempenho da detecção de pedestres, foi utilizado o detector *VeryFast* combinado com *stixels*, cujos códigos estão disponibilizados *online* por Benenson et al. (2012b) e Benenson et al. (2012a).

3 REVISÃO BIBLIOGRÁFICA

As duas principais abordagens adotadas para detecção de pedestres são as baseadas em janelas deslizantes e as que utilizam arquiteturas profundas de redes neurais (DOLLÁR et al., 2012; GEIGER; LENZ; URTASUN, 2012). Atualmente, os métodos que possuem melhor desempenho são baseadas em arquiteturas profundas de redes neurais, porém, seus custos computacionais são elevados, impedindo a execução em tempo real. Em contrapartida, os métodos com abordagem em janelas deslizantes são rápidos e mantêm o bom desempenho. Portanto, na Seção 3.1 é realizada uma revisão dos métodos existentes de detecção de pedestres com o foco na abordagem em janelas deslizantes, e na Seção 3.2 são apresentados os principais *benchmarks* em detecção de pedestres, de modo a evidenciar as razões para a configuração de câmeras proposta. Neste capítulo e nos demais, o termo desempenho refere-se somente à taxa de acerto dos métodos de detecção de pedestre sem considerar o seu custo computacional, que por sua vez tem influência no tempo de processamento.

3.1 Trabalhos Relacionados

Papageorgiou e Poggio (2000) popularizaram o uso de detectores com abordagem em janelas deslizantes após proporem um dos primeiros detectores de objetos baseado em tal arquitetura. Os autores abordaram a aprendizagem baseada em exemplos, que resulta em um modelo através do treinamento da Máquinas de Vetores Suporte (SVM) com a transformada de *Haar wavelet* como característica. Após isso, vários trabalhos foram desenvolvidos com base neste tipo de arquitetura, variando apenas os descritores de características e algoritmos de aprendizagem (classificadores).

Inicialmente, os trabalhos apenas exploravam as características individualmente, tais como *Haar wavelet* (PAPAGEORGIOU; POGGIO, 2000), *Haar-like* (VIOLA; JONES, 2004), Histograma de Gradientes Orientados (DALAL; TRIGGS, 2005) e *Shape Context* (LEIBE; SEEMANN; SCHIELE, 2005). Porém, em uma avaliação realizada por Wojek e Schiele (2008), os autores observaram que a combinação de diferentes características com diferentes classificadores apresentam variações no desempenho, e apenas os descritores HOG e *Shape Context* mantiveram os melhores desempenho independentemente do algoritmo de aprendizagem utilizado. Dessa forma, os autores propuseram um detector de pedestres (MultiFtr) que utiliza a combinação de várias características (*Haar wavelet*, HOG e *Shape Context*), e mostraram que a combinação de características supera qualquer característica individual, inclusive o método estado-da-arte da época publicado por Dalal e Triggs (2005). Desde então, foram desenvolvidos vários trabalhos que propõe novas combinações de características com técnicas de imagens integrais para alcançar melhor desempenho

(DOLLÁR et al., 2009; WANG; HAN; YAN, 2009; DOLLÁR; BELONGIE; PERONA, 2010; BENENSON et al., 2012b; DOLLÁR et al., 2014; NAM; DOLLÁR; HAN, 2014).

Em Dollár et al. (2009), foi estudado o uso de canais integrais de características para detecção de pedestres, com a ideia de que uma representação de canais de imagens pode ser computada utilizando transformações lineares e não-lineares de uma imagem de entrada, e posteriormente, através do uso de imagens integrais, as características são extraídas para cada canal usando a soma local de regiões retangulares. Por meio de uma avaliação da combinação de diferentes tipos de canais (histograma de gradientes, cor e magnitude de gradiente), observou-se que o canal mais informativo é o de orientação de gradientes. Porém, o desempenho satura quando é utilizado mais que seis orientações de gradientes. Dessa forma, foi proposta pelos autores a combinação da magnitude de gradiente, histograma de seis orientações de gradientes e canais de cor LUV, dando origem ao detector *Integral Chanel Features* (ICF).

Os métodos tradicionais em detecção de pedestres necessitam que a imagem de entrada seja redimensionada N vezes para que seja possível realizar a detecção em multi-escalas, e então as características são computadas para cada uma dessas imagens redimensionadas. Porém, Viola e Jones (2004) mostraram que é mais vantajoso redimensionar as características ao invés das imagens. No entanto, era possível realizar esse redimensionamento somente em características invariantes a translação e escala.

Inspirado na ideia de Viola e Jones (2004), Dollár, Belongie e Perona (2010) abordaram em seu trabalho o gargalo computacional dos detectores, que é a necessidade de computar as características para cada escala da pirâmide de imagem (aproximadamente 8 a 16 escalas por oitava). Com essa motivação, foi proposta pelos autores uma técnica que permite a computação aproximada de características em escalas adjacentes, a partir de características computadas apenas para uma escala. Através de experimentos realizados, observou-se que a qualidade de aproximação degrada com o aumento do intervalo entre as escalas. Dessa forma, foi proposto um método híbrido que constrói amostras de pirâmide de imagem com intervalos de uma oitava, e dentro de cada oitava é utilizado a pirâmide de classificadores. Com base nas características utilizadas pelo detector *ChnFtrs* (DOLLÁR et al., 2009), esse trabalho resultou em um novo método para detecção de pedestres, denominado *Fastest Pedestrian Detector in the West* (FPDW), obtendo desempenho bem próximo do método original. Porém, sendo aproximadamente entre 5 a 10 vezes mais rápido que o *ChnFtrs*, com taxa de execução variando de 2,67 a 6,49 Hz.

Até então, houve uma grande melhora no desempenho das detecções. Porém, os métodos que garantiam melhor desempenho normalmente vinham acompanhado de um alto custo computacional. Dessa forma, alguns autores buscaram acelerar a taxa de execução através de melhora nos métodos de computar características, algoritmos de classificação, uso de informações de profundidade, etc. (BENENSON et al., 2012b; BENENSON et al.,

2012a; DOLLÁR et al., 2014).

Em Benenson et al. (2012b), os autores apresentaram um novo detector de pedestres (*VeryFast*) que melhorou tanto a velocidade quanto a qualidade sobre os métodos estado-da-arte da época. A ideia dos autores foi realizar o inverso da proposta do detector FPDW, utilizando apenas uma escala de imagem de entrada e redimensionando o modelo treinado. Isto é, N/K (~ 5) modelos foram treinados para diferentes escalas, e então os N/K modelos foram redimensionados em N (~ 50) modelos através de uma aproximação. Na avaliação realizada pelos autores, ao realizar o redimensionamento na etapa de treinamento ao invés da etapa de teste, o ganho de velocidade proporcionado é maior que 7 vezes em comparação ao método FPDW, resultando em detecções a 50 *Hz* no INRIA *dataset*. Para melhorar a qualidade e a velocidade da detecção, os autores utilizaram uma variação do *stixel world* que permite reduzir a região de procura por um fator de 44, anteriormente proposto em Benenson, Timofte e Van Gool (2011). Este método consiste em estimar rapidamente o *pixel* inferior, superior e a distância do objeto diretamente a partir das imagens estéreo, sem a necessidade de computar completamente o mapa de profundidade. Quando é realizado a combinação do *VeryFast* com as informações de profundidade (*stixels*), o detector executa a uma taxa de 135 *Hz*.

Dollár et al. (2014) propuseram um método rápido e eficiente de calcular pirâmides de características sem perder o desempenho. A fim de demonstrar a eficácia do *fast feature pyramids*, os autores utilizaram uma variação do *Integral Channel Features*, denominado *Aggregated Channel Features* (ACF), que também utiliza como característica canais de magnitude de gradiente normalizado, histograma de gradientes orientados (6 orientações) e canais de cor LUV. A comparação de desempenho do ACF foi realizada com 16 detectores no INRIA *Pedestrian Dataset*, com imagens de 640×480 *pixels*. O método proposto pelos autores executou a aproximadamente 32 *fps* com uma taxa de erro de 17%. Porém, nessa avaliação, o detector *VeryFast* proposto em Benenson et al. (2012b), mostrou ser mais interessante ao alcançar 16% de taxa de erro com velocidade de 50 *Hz*.

Devido sua simplicidade e inferência de velocidade, os classificadores lineares normalmente são utilizados em métodos de janelas deslizantes (BENENSON et al., 2012b). No entanto, os recentes avanços em unidades de processamento gráfico (GPU) e em técnicas de aprendizado profundo têm permitido que redes neurais com multi-camadas atinjam menores taxas de erro em qualquer aplicação de visão computacional, superando em termos de acerto, os métodos tradicionais com abordagem em janelas deslizante, tais como R-CNN (GIRSHICK et al., 2014), MS-CNN (CAI et al., 2016), SA-FastRCNN (LI et al., 2015) e F-DNN+SS (DU et al., 2016). Contudo, este grande ganho vem acompanhado de um alto custo computacional.

3.2 Desempenho dos detectores

Durante os últimos anos, vários trabalhos têm sido desenvolvidos para o problema de detecção de pedestres. Mas, devido ao uso de diferentes *datasets*, variações no protocolo de avaliação e complexidade das diferentes propostas, se torna difícil realizar a comparação direta entre os métodos, e a maioria dos autores acabam apresentando apenas a sua própria abordagem. Dessa forma, com o propósito de comparação e avaliação, diversos *benchmarks* foram desenvolvidos para detecção de pedestres (DALAL; TRIGGS, 2005; ENZWEILER; GAVRILA, 2009; DOLLÁR et al., 2009; DOLLÁR et al., 2012; GEIGER; LENZ; URTASUN, 2012; BENENSON et al., 2015; ZHANG; BENENSON; SCHIELE, 2017).

Em Dollár et al. (2009), foi realizado um *benchmark* dos detectores mais promissores, fornecendo uma visão geral do desempenho estado-da-arte e uma comparação direta e justa dos métodos. Para isto, foi introduzido um novo conjunto de dados, o *Caltech Pedestrian Dataset*, no qual os pedestres possuem uma ampla variação na aparência, pose e escala, além de possuir informações de oclusão anotadas. Através da avaliação, os autores observaram que até o melhor método avaliado (MultiFtr) não tinha um bom desempenho absoluto, com taxa de erro acima de 80%. Dessa forma, para entender onde os métodos falham, os autores analisaram o desempenho dos detectores variando apenas subconjuntos dos dados, agrupando os pedestres de acordo com sua razão de aspecto, escala e oclusão.

As significantes variações entre os *bounding boxes* (bbox) do *ground truth* e da detecção realizada causa um efeito indesejado na avaliação, pois, enquanto a altura dos bbox de pedestres dependem apenas da sua escala, a largura depende também da sua pose. Com isso, tanto a largura do bbox do *ground truth* quanto da detecção são mais variáveis e aleatórios do que a altura dos bbox. Então, para remover qualquer efeito que isso possa ter sobre a avaliação de desempenho, os autores propuseram padronizar todos os bbox do *ground truth* para uma razão média de aspecto de 0,43 ($\frac{width}{height} = 0,43$), com um desvio padrão de 0,1. Dessa forma, os bbox que estiverem fora desse padrão são considerados na avaliação como razão de aspecto atípica.

Em relação ao agrupamento dos pedestres de acordo com sua altura em *pixel*, essa divisão foi realizada para pedestres com escala grande (maior que 80 *pixels*), média (30 ~ 80 *pixels*) e pequena (menor que 30 *pixels*). Por fim, Dollár et al. (2009) agruparam os pedestres com altura mínima de 50 *pixels* de acordo com o grau de oclusão dos mesmos, sendo agrupados como sem oclusão, oclusão parcial (1 ~ 35% da área oclusa) e oclusão densa (35 ~ 80%) da área oclusa. Como resultado, foi possível observar que as maiores dificuldades são pedestres em menores escalas e sob efeito de oclusão, onde, o desempenho cai drasticamente em situações de pequena escala (92% de taxa de erro obtido pelo melhor método avaliado) e oclusão densa (93% de taxa de erro).

Figura 17: Resultados dos algoritmos avaliados pelo *Caltech Pedestrian Detection Benchmark* no *dataset* da Caltech, com atualização realizada em Dezembro de 2016. (a) Desempenho absoluto realizado em todo o *dataset*. (b-c) Razão de aspecto típica e atípica. (d) Pedestres maiores que 80 *pixels*. (e) Pedestres com altura entre 30 a 80 *pixels*. (f) Pedestres com altura menor que 30 *pixels*. (g) Pedestre sem oclusão. (h) Pedestres com oclusão parcial (1 ~ 35% da área oclusa). (i) Pedestres com oclusão densa (35 a 80% da área oclusa).

Fonte: Dollár et al. (2009).

Posteriormente, Dollár et al. (2012) realizaram uma nova avaliação dos métodos estado-da-arte em diversos *datasets*, incluindo ETH (ESS; LEIBE; Van Gool, 2007), TUD-Brussels (WOJEK; WALK; SCHIELE, 2009), Daimler (ENZWEILER; GAVRILA, 2009), INRIA (DALAL; TRIGGS, 2005) e duas variações do Caltech, mostrando que o

desempenho de um detector relativo aos outros independe do *dataset* utilizado. Isto é, a classificação do detector em um *benchmark* será razoavelmente consistente em qualquer *dataset* testado.

Atualmente, os principais *benchmarks* em detecção de pedestres são o KITTI *Vision Benchmark Suite* (GEIGER; LENZ; URTASUN, 2012) e o *Caltech Pedestrian Detection Benchmark* (DOLLÁR et al., 2009), que estão disponíveis *online* para os autores enviarem seus trabalhos para serem avaliados. Em ambos *benchmarks*, os métodos que estão mostrando ser mais promissores em termos de desempenho (*recall* e *precision*) são os que utilizam arquiteturas profundas de redes neurais, tais como Li et al. (2015), Du et al. (2016), Cai et al. (2016), Ren et al. (2017).

Em comparação ao KITTI, o benchmark da Caltech é mais desafiador e completo. Além disso, a avaliação é realizada da mesma forma como foi proposto em Dollár et al. (2009), agrupando os pedestres de acordo com sua escala, razão de aspecto e oclusão, o que permite uma análise mais detalhada do desempenho de cada método. A Figura 17 apresenta os resultados da comparação de desempenho entre os métodos no *benchmark* da Caltech. Cada gráfico é composto apenas pelos 15 métodos que obtiveram melhor resultado na avaliação, e os valores nas legendas representam a taxa de erro ($MR = 1 - recall$) de cada método para suas respectivas condições, ou seja, a porcentagem de pedestres que os detectores não conseguiram identificar. Através da Figura 17(a), pode-se observar que o resultado absoluto destes métodos ainda é precário, pois, pelo menos metade dos pedestres não são detectados. E, o maior desafio consiste em detectar pedestres em escalas menores, ou seja, pelo menos 77% dos pedestres que estão mais distantes das câmeras não são detectados (Figuras 17(e) e 17(f)). Assim como os pedestres que estão com 35 a 80% de seus corpos sob efeito de oclusão (Figura 17(i)).

Apesar da relevante evolução de desempenho alcançada durante os últimos anos, ainda têm-se muito a melhorar. Recentemente, Zhang et al. (2016) mostraram que os melhores detectores são inferiores, em todos os aspectos, aos humanos. Além disso, esses métodos que obtêm melhor desempenho (F-DNN+SS, RPN+BF, SA-FastRCNN, MS-CNN, CompACT-Deep, etc.) são acompanhados de um alto custo computacional, proporcionando taxas de detecções mensuráveis em segundos por *frame*, e, a execução em tempo real é crucial para aplicações automobilísticas. De acordo com o *Caltech Pedestrian Detection Benchmark*, o atual método estado-da-arte em detecção de pedestres é baseado em uma fusão de arquiteturas profundas de redes neurais convolutivas (F-DNN+SS). No geral, o desempenho deste método é muito melhor que os métodos tradicionais baseados em janelas deslizantes, porém, seu tempo de processamento é muito lento (2,48 segundos por *frame* executando em uma NVIDIA Titan X).

3.3 Considerações finais

Neste capítulo foi realizada uma revisão dos métodos visuais de detecção de pedestres. Como pôde ser observado na Seção 3.2, apesar da considerável melhora no desempenho alcançado pelos recentes métodos, os detectores ainda apresentam grande dificuldade em situações de pedestres em menores escalas, onde o melhor método avaliado (F-DNN+SS) deixa de detectar 77% dos pedestres contidos no *dataset* da Caltech. Porém, esses pedestres que estão mais distantes do veículo também devem ser detectados para que seja possível evitar uma colisão. Ademais, as aplicações automobilísticas exigem que as detecções sejam realizadas em tempo real para que haja uma possibilidade do veículo conseguir frear a tempo, enquanto isso, o tempo de execução do F-DNN+SS é muito lento (em torno de 0,4 *fps* em um NVIDIA Titan X), mostrando que com a tecnologia atual ainda é inviável usar métodos com arquiteturas profundas de redes neurais.

Apesar do detector *VeryFast* não possuir um desempenho tão bom quanto os métodos mais recentes, ele apresenta grandes vantagens para aplicações automobilísticas devido seu rápido tempo de execução. No entanto, sua desvantagem consiste em detectar pedestres que estão distantes do veículo devido suas pequenas escalas nas imagens. Com isso, se houvesse uma forma de superar o desafio em relação a detecção de pedestres em menores escalas, o uso do detector *VeryFast* seria uma melhor alternativa.

4 APLICABILIDADE DOS MÉTODOS

Os *benchmarks* existentes em detecção de pedestres possibilitaram verificar a evolução dos métodos propostos durante os últimos anos. Apesar do significativo progresso alcançado pelos recentes métodos, ainda não é possível saber se o desempenho estado-da-arte é o suficiente para aplicações automobilísticas. Dessa forma, com o propósito de analisar a aplicabilidade dos métodos de detecção de pedestres para assistir o motorista e evitar eventuais acidentes, foi realizado um estudo para verificar a qual distância o pedestre deve ser detectado para que o veículo consiga frear a tempo. Os resultados obtidos neste capítulo foram publicados em Nakamura, Horita e Grassi Jr. (2017a).

4.1 Metodologia

O estudo foi realizado considerando a dinâmica do veículo durante a frenagem, as limitações atuais de *hardware* e *software* que o sistema possui, e a altura dos pedestres na imagem. A distância necessária para o veículo ser capaz de frear a tempo é denominada de distância total de frenagem, definida como:

$$D_t = D_i + D_r + D_s + D_f \quad (4.1)$$

onde, D_i é a distância total percorrida pelo veículo durante o tempo de aquisição da imagem, D_r é a distância percorrida correspondente ao tempo de execução do sistema de detecção de pedestres, D_s corresponde a distância devido ao tempo de atuação efetiva do freio do veículo após o acionamento do pedal, e D_f é a distância percorrida após o acionamento do freio até a parada do veículo.

Figura 18: Distância entre o sistema de visão e a frente do veículo.

Fonte: Elaborado pela autora.

Dado que a câmera normalmente fica instalada na parte superior central do veículo, deve ser somado a D_t , a distância x entre a frente do carro e o sistema de visão (Figura

18). Dessa forma, a mínima distância que o pedestre deve ser detectado consiste em, $D_{tf} = D_t + x$. Os parâmetros utilizados para calcular D_t serão apresentados a seguir.

4.1.1 Distância de frenagem

Considerando um veículo com velocidade inicial V_1 , a distância de frenagem pode ser obtida através da seguinte equação (WONG, 2008):

$$D_f = \frac{\gamma W}{2gC_{ae}} \ln \left(1 + \frac{C_{ae} V_1^2}{\eta \mu W + f_r W \cos \theta \pm W \sin \theta} \right), \quad (4.2)$$

onde:

$$C_{ae} = (\rho A_f C_d)/2;$$

$$\gamma = \text{Fator equivalente de massa};$$

$$g = \text{Aceleração da gravidade } (m/s^2);$$

$$\rho = \text{Massa específica do ar } (kg/m^3);$$

$$\eta = \text{Eficiência do freio};$$

$$\mu = \text{Coeficiente de atrito};$$

$$f_r = \text{Coeficiente de resistência do rolamento};$$

$$W = \text{Peso do veículo } (kg);$$

$$A_f = \text{Área frontal do veículo } (m^2);$$

$$C_d = \text{Coeficiente de aerodinâmica};$$

$$\theta = \text{Ângulo de inclinação da pista};$$

$$V_1 = \text{Velocidade inicial do veículo } (m/s).$$

4.1.2 Distância de aquisição de imagem

A distância percorrida durante o tempo de aquisição da imagem é calculada como sendo $D_i = t_i V_1$, onde, t_i é o intervalo de tempo entre a captura de quadros de imagem consecutivos.

4.1.3 Distância de reação

O tempo de reação consiste no tempo que o sistema de detecção ou o condutor necessita para identificar um pedestre a sua frente. A distância que o veículo percorre durante este tempo é chamado de distância de reação, dado por $D_r = t_r V_1$. Onde, t_r representa o tempo de processamento do método de detecção de pedestre após a captura da imagem.

4.1.4 Distância de atuação do sistema de freio

O tempo de atuação do sistema de freio, t_s , é o tempo de resposta do mesmo a partir do momento em que o pedal do freio é acionado. Desse modo, $D_s = t_s V_1$ é a

distância percorrida pelo veículo durante o início da ação até a configuração completa da pressão de frenagem. Segundo (CHEN et al., 2013), o tempo de atuação do sistema de freio normalmente varia entre 0,3 e 0,75 segundos.

4.2 Estudo de caso

Para o estudo de caso, considerou-se a plataforma CaRINA II (Carro Robótico Inteligente para Navegação Autônoma). A plataforma CaRINA II é um Fiat Palio Adventure modificado pelo Laboratório de Robótica Móvel (LRM) no Instituto de Ciências Matemáticas e de Computação (ICMC-USP), para possibilitar operações autônomas através do controle completo de suas funções (aceleração, freio e esterçamento) por meio de algoritmos de percepção e controle (SHINZATO et al., 2016).

Tendo em vista o veículo Fiat Palio Adventure Dualogic com *Anti-Lock Braking System* (ABS), foram utilizados os parâmetros apresentados na Tabela 1 para calcular a distância de frenagem (D_f). Os valores dos parâmetros foram obtidos a partir das configurações do veículo e dos trabalhos de Wong (2008) e Gardinalli (2005).

Tabela 1: Parâmetros e valores usados para o cálculo da distância de frenagem.

Parâmetro		Configuração
γ	Fator equivalente de massa	1,04
g	Aceleração da gravidade	9,81 m/s^2
ρ	Massa específica do ar	1,30 kg/m^3
η	Eficiência do freio	0,90
μ	Coeficiente de atrito	Pista seca 0,85 Pista molhada 0,60
f_r	Coeficiente de resistência do rolamento	0,0015
W	Peso do Veículo Palio Adventure (1.198 kg) + 2 pessoas (140 kg)	1.338 kg
A_f	Área frontal do veículo	2,60 m^2
C_d	Coeficiente de aerodinâmica	0,35
θ	Ângulo de inclinação da pista	0°

Fonte: Elaborado pela autora.

A plataforma CaRINA II utiliza a câmera *Bumblebee XB3* da Point Grey (modelo BBX3-13S2C-38, com resolução de 1280×960 , 3,8 mm de distância focal, tamanho do *pixel* de 3,75 μm e 24 cm de *baseline*) para coletar as imagens estéreo a uma taxa 16 *fps*, ou seja, $t_i = 0,06$ s.

Apesar de não possuir uma taxa de acerto tão boa quanto outros métodos de detecção de pedestres mais recentes, o método *VeryFast* (BENENSON et al., 2012b) foi selecionado para esta avaliação devido ao seu rápido tempo de execução, que varia de 50 a 135 *Hz*. Para a mesma taxa de acerto do *VeryFast*, os demais métodos não conseguem

Figura 19: Estimativas da distância total percorrida por um veículo em situações de frenagens bruscas, considerando o tempo de reação do detector *VeryFast* (50 *Hz*) em condições de pista seca e molhada.

Fonte: Elaborado pela autora e publicado em Nakamura, Horita e Grassi Jr. (2017a).

fazer a execução tão rapidamente. Dessa forma, para realizar a análise foi adotado o tempo de execução de 50 *Hz*, ou seja, $t_r = 0,02$ s.

Com todos os valores dos parâmetros definidos, é possível estimar através da Equação 4.2, a distância total percorrida pelo veículo em situações de frenagens bruscas. Os resultados dos cálculos são apresentados na Figura 19. Como pode ser observado na Figura 19(a), um veículo a 60 *km/h* em uma pista seca percorre 26 metros até conseguir parar completamente. E, em condições de pista molhada, essa distância aumenta para 32 metros (Figura 19(b)). No CaRINA II, a câmera fica instalada a 2 metros de distância da parte frontal do veículo. Desse modo, considerando $D_t = 32$ metros, o sistema deve identificar pedestres que estejam pelo menos a 34 metros de distância das câmeras.

Figura 20: Formação de imagem em uma máquina fotográfica.

Fonte: Elaborado pela autora.

Figura 21: Distância vs. altura do pedestre na imagem (*Bumblebee XB3*).

Fonte: Elaborado pela autora.

Definido a distância mínima, é possível obter a altura que os pedestres terão na imagem através do modelo de câmera *pinhole*. E, dessa forma, estimar a menor escala de pedestres que os métodos deverão ser capazes de detectar. Esta altura pode ser obtida através da relação entre a altura do pedestre (h_o), a altura do pedestre na imagem (h_i), a distância do pedestre à câmera (d) e a distância focal da câmera (f) (Figura 20).

Conhecendo a distância do pedestre em relação à câmera (d), a altura do pedestre (h_o), a distância focal (f) e o tamanho do *pixel* ($3,75 \mu m$), calcula-se a altura do pedestre na imagem. Os cálculos foram realizados considerando pedestres com altura variando entre 1 a 2 metros. Então, de acordo com a Figura 21, os métodos devem detectar pedestres com altura na imagem maior ou igual aos mostrados na Tabela 2.

Tabela 2: Relação entre a distância mínima que os pedestres devem ser detectados e sua altura na imagem para que seja possível realizar a frenagem a tempo.

Distância da câmera (m)	Altura do Pedestre (m)	Altura na imagem ($pixels$)
34	1,0	29
	1,5	45
	2,0	59

Fonte: Elaborado pela autora.

Entretanto, como foi mostrado na Seção 3.2, em situações de pedestres menores que 30 *pixels*, mais de 77% dos pedestres contidos no *dataset* não são detectados, enquanto que para escalas entre 30 a 80 *pixels*, a taxa de erro reduz para 33%. Apesar desta significativa melhora no desempenho, supõe-se que a taxa de erro não esteja igualmente distribuída para todo o intervalo da escala média. Assim como há um declínio de desempenho da escala média para a pequena, assume-se que dentro de um intervalo de escala haja uma queda contínua de desempenho com a redução de escala. Isto é, os detectores irão apresentar maior dificuldade quanto menor for a escala do pedestre.

Em comparação aos humanos, a aplicação de sistemas de detecção de pedestres em veículos apresenta uma grande vantagem em relação ao tempo de reação. De acordo com o trabalho de Johansson e Rumar (1971), o tempo médio de reação de uma pessoa adulta, em situações de frenagens inesperadas, é de aproximadamente 0,9 segundos. Apenas neste intervalo de tempo, a diferença de distância percorrida entre um veículo conduzido por um motorista e outro com um sistema de detecção de pedestres é muito significativo. Essa diferença pode ser observada na Figura 22, onde a distância percorrida por um veículo a 60 *km/h* durante o tempo que o sistema leva para detectar o pedestre é 45 vezes menor do que a distância percorrida durante o tempo que um humano levaria para reagir.

Figura 22: Comparação da distância percorrida durante o tempo de reação de um ser humano e de um sistema de detecção de pedestres.

Fonte: Elabora pela autora.

4.3 Considerações finais

Neste capítulo foi realizada uma análise da aplicabilidade dos métodos estado-da-arte de detecção de pedestres para veículos inteligentes, considerando as limitações atuais de *hardware* e *software* que o sistema possui, o contexto real de cenários urbanos, a dinâmica e velocidade do veículo durante a frenagem, e a altura dos pedestres na

imagem. Para isso, as características físicas da plataforma CaRINA II foram tomadas como referência do estudo.

Através do estudo realizado, conclui-se que para as condições determinadas neste trabalho, o pedestre deve ser detectado a pelo menos 34 metros de distância da câmera, onde sua altura na imagem corresponde a apenas 29 *pixels* (considerando a câmera *Bumblebee XB3*). Apesar da última atualização do *Caltech Pedestrian Detection Benchmark* (DOLLÁR et al., 2009) ter mostrado que os recentes métodos estão conseguindo melhorar o desempenho em relação a pedestres em pequenas escalas, observa-se que, ao considerar apenas o desempenho obtido pelos métodos estado-da-arte, seria possível utilizar detectores de pedestres para assistir o motorista somente nas situações em que os pedestres estiverem em escalas maiores (região superior da Figura 21), onde a taxa de erro é nula. Em outras palavras, a operação do sistema estaria limitado a no máximo 12 metros à frente da câmera. Assim, de acordo com a Figura 19(b), o veículo teria que estar abaixo de 30 *km/h* para evitar a colisão.

No entanto, é importante recordar que a melhora dos recentes métodos veio acompanhado de um alto custo computacional com execução a aproximadamente 0,4 *fps*, uma vez que eles utilizam arquiteturas profundas de redes neurais. Apesar da constante evolução na capacidade computacional, outras tarefas de mais alto nível são executadas nos veículos inteligentes, tais como localização, mapeamento, detecção de área navegável, etc. Porém, ao realizar uma análise detalhada do desempenho dos detectores no *benchmark* da Caltech, em situações de pedestres em maiores escalas (Figura 17(d)), a diferença de desempenho entre os métodos que utilizam arquiteturas profundas de redes neurais e os métodos com abordagem em janelas deslizantes não é significativa, enquanto que a diferença no tempo de execução é muito grande. Assim, se houvesse uma forma de adquirir imagens com pedestres distantes das câmeras em maiores escalas, não seria necessário aumentar a capacidade computacional e utilizar métodos com arquiteturas profundas.

5 PROPOSTA DE CONFIGURAÇÃO DE CÂMERAS ESTÉREO

Como forma de melhorar o desempenho da detecção visual de pedestres, propõe-se utilizar uma configuração de câmeras que possibilitará a aquisição de imagens com pedestres em maiores escalas e em um maior intervalo de distância à frente do veículo. Isto será feito através da combinação de dois pares de câmera estéreo que trabalharão com diferentes distâncias focais e *baselines*. Os resultados obtidos neste capítulo foram publicados em Nakamura, Horita e Grassi Jr. (2017b). Observa-se que o termo câmera estéreo utilizado neste trabalho refere-se a um sistema composto por duas câmeras monoculares.

Figura 23: Sistema de câmeras proposta instalada na plataforma CaRINA II.

Fonte: Elaborado pela autora e publicado em Nakamura, Horita e Grassi Jr. (2017b).

5.1 Visão geral do sistema

Para implementar o sistema de visão estéreo, quatro câmeras Flea3 da Point Grey (modelo U3-20E4C-C, com resolução de 1600×1200 , formato do sensor de $1/1,8''$ e tamanho do *pixel* de $4,5 \mu m$) foram alinhadas horizontalmente em uma barra que posteriormente foi instalada no CaRINA II (Figura 23). Nesta configuração, cada câmera estéreo é responsável por operar em diferentes intervalos de distância à frente do veículo. Em geral, como as *baselines* maiores permitem estimativa de profundidade com melhor acurácia em longas distâncias (OKUTOMI; KANADE, 1993), cada *baseline* foi selecionada de forma a ter maior distância entre as câmeras e operar na região definida. O esquema da configuração desse sistema de pares estéreo é apresentado na Figura 24, onde as câmeras internas formam o sistema estéreo (62 cm de baseline) para detectar pedestres que estão a uma

Figura 24: Configuração do sistema de câmeras proposto para detecção de pedestres (dimensões em metros).

Fonte: Elaborado pela autora.

distância de 5 a 20 *m* à frente das câmeras, enquanto que as câmeras externas formam o sistema estéreo (93 *cm* de *baseline*) para detectar pedestres que estão a uma distância maior que 20 *m*.

Para as câmeras internas, foram utilizadas duas lentes (Edmund Optics) com 4,5 *mm* de distância focal, 84,7° de ângulo de visão (AOV) e formato do sensor de 1/1,8", possibilitando uma visão mais ampla da região próxima ao veículo. Diferentemente, a

Figura 25: Diagrama de operação para cada câmera estéreo do sistema proposto.

Fonte: Elaborado pela autora.

fim de adquirir imagens com pedestres distantes das câmeras e em maiores escalas, o sistema estéreo externo utilizou lentes com $8,5\text{ mm}$ de distância focal (Edmund Optics), $59,1^\circ$ de AOV e formato do sensor de $2/3''$. A Figura 25 fornece uma visão geral do funcionamento do sistema proposto para cada par de câmeras. Primeiro, é realizada a calibração das câmeras para obter os parâmetros das câmeras e posteriormente, retificar as imagens e recuperar as informações de profundidade a partir das imagens adquiridas. Com isso, o *ground plane* e os *stixels* são computados para cada câmera estéreo, retornando a localização dos obstáculos. Dessa forma, o detector de pedestres *VeryFast* é executado somente na região onde foi considerado a existência de obstáculos, e, caso exista pedestres nessa região, é retornado os *bounding boxes* dos pedestres identificados em cada sistema estéreo.

Em comparação à câmera *Bumblebee XB3* (Figura 21), o sistema proposto será capaz de realizar a aquisição de imagens com pedestres em maiores escalas (aproximadamente acima de 80 pixels) e em um intervalo de distância mais amplo (Figura 26).

Figura 26: Distância vs. altura do pedestre na imagem em relação a diferentes distâncias focais. (a) Distância focal de $4,5\text{ mm}$. (b) Distância focal de $8,5\text{ mm}$.

Fonte: Elaborado pela autora.

5.1.1 Distância focal equivalente

Como o formato do sensor das câmeras utilizadas ($1/1,8''$) é menor que o das lentes com distância focal de $8,5\text{ mm}$ ($2/3''$), a área capturada pelo sensor é menor que a área fornecida pelas lentes. Dessa forma, haverá um efeito de zoom causado pelo corte nas bordas da imagem fornecida pela lente, o qual pode ser interpretado como uma imagem capturada com distância focal maior (Figura 27). Essa distância focal é conhecida como distância focal equivalente (f_{eq}), expressa por:

$$f_{eq} = f \cdot \frac{d_{lente}}{d_{sensor}},$$

onde, f é a distância focal da lente, d_{lente} é o tamanho diagonal do formato do sensor fornecido pela lente, e d_{sensor} é o tamanho diagonal do sensor da câmera. Para os sensores com formato de $1/1,8''$ e $2/3''$, os tamanhos diagonais são $8,93\text{ mm}$ e $11,00\text{ mm}$, respectivamente (Digital Photography Review, 2002). Dessa forma, para o par de câmeras com lentes de $8,5\text{ mm}$, a distância focal equivalente será de aproximadamente $10,54\text{ mm}$. Em relação às câmeras com lentes de $4,5\text{ mm}$, não haverá alteração na distância focal pois não há diferença de tamanho entre o formato do sensor das câmeras e das lentes.

Figura 27: Efeito de corte causado pela diferença de tamanho entre a lente e o sensor.

Fonte: Elaborado pela autora.

5.1.2 Região de interesse

As câmeras utilizadas neste trabalho fornecem imagens em alta resolução ($1600 \times 1200\text{ pixels}$), porém, a aquisição das imagens foi limitada para uma região de interesse (ROI) menor. Essa técnica possibilita que apenas uma região específica da imagem seja transmitida pela câmera, proporcionando o envio de imagens em maior *frame rate*.

Dessa forma, cada câmera estéreo teve a ROI de $900 \times 600\text{ pixels}$ limitada em uma posição diferente, conforme apresentada na Figura 28. Para definir a posição de cada ROI, foi considerada a região mais propícia de haver uma colisão entre o veículo e os pedestres (lado da pista que o veículo está percorrendo), e o intervalo de distância que cada câmera irá operar. Dado que o par de câmeras com *baseline* maior ($8,5\text{ mm}$ de distância focal) é utilizado para lidar com detecção de pedestres que estão mais distantes do veículo, a ROI foi posicionada em uma região superior da imagem (Figura 28(b)) se comparada à posição da ROI definida para as imagens obtidas pelo outro par de câmeras (Figura 28(a)).

Figura 28: Diferentes regiões de interesse utilizadas para cada sistema estéreo.

(a) $f = 4,5 \text{ mm}$

(b) $f = 8,5 \text{ mm}$

Fonte: Elaborado pela autora.

Ademais, ao limitar a região de operação para cada imagem, a quantidade de dados a serem processados pelo detector de pedestres é reduzida, permitindo melhor eficiência.

Quando as quatro câmeras Flea3 são sincronizadas para capturar imagens simultaneamente, a taxa de aquisição atinge até 77 fps com o ROI limitado para uma região de $900 \times 600 \text{ pixels}$.

5.2 Calibração das câmeras

O procedimento de estimar os valores dos parâmetros intrínsecos e extrínsecos da câmera é denominado de calibração. Segundo Trucco e Verri (1998), a ideia da calibração consiste em definir equações de projeção relacionando as coordenadas conhecidas de um conjunto de pontos 3D e suas projeções para estimar os parâmetros da câmera. Para isso, os autores apontam que os métodos de calibração dependem de um padrão com geometria conhecida, que é localizado em uma posição conhecida no espaço para gerar características

Figura 29: Calibração das câmeras

(a) Detecção do padrão em todo campo de visão superior, inferior, esquerda e direita.

(b) Detecção em vários ângulos.

(c) Detecção em diversos tamanhos.

(d) Padrão inclinado para a esquerda, direita, em cima e embaixo.

Fonte: Elaborado pela autora.

nas imagens que possam ser localizados com precisão. Porém, Zhang (2000) desenvolveu uma técnica que necessita apenas de uma câmera observando um padrão planar em diferentes orientações, onde, tanto a câmera quanto o padrão podem ser movimentados manualmente. Por consistir de uma técnica mais simples e flexível se comparada com os métodos clássicos, muitos trabalhos utilizam o método de Zhang (2000) para realizar a calibração das câmeras, inclusive a biblioteca OpenCV, que é utilizada no ROS. Dessa forma, para modelar as distorções das lentes e encontrar os parâmetros intrínsecos e extrínsecos, foi utilizado o pacote *camera_calibration* do ROS, e um padrão xadrez 8×6 com quadrados de 108 mm .

Devido ao padrão utilizado ser uma placa maleável, o mesmo foi fixado na parede para evitar imprecisão devido às distorções no padrão (Figura 29). Com isso, as câmeras foram movidas em volta ao padrão de forma a detectá-lo em todo o campo de visão do eixo x e y da imagem (Figura 29(a)), em diferentes distâncias das câmeras (Figura 29(c)), e em vários ângulos de inclinação e rotação (Figuras 29(b) e 29(d)). Com o resultado da calibração, foi possível retificar as imagens e recuperar as informações de profundidade (Figura 30), que posteriormente foram utilizados no algoritmo *VeryFast*.

Figura 30: Resultado da calibração. (a-b) Imagens esquerda e direita retificadas. (c) Mapa de disparidade.

Fonte: Elaborado pela autora.

5.3 Dataset

O principal objetivo da configuração de câmeras proposta é superar o baixo desempenho na detecção de pedestres em pequenas escalas. Então, para avaliar a eficácia do sistema proposto, foi coletado um novo conjunto de dados com cinco pedestres entre 1,5 a 1,8 metros de altura caminhando em frente ao veículo a uma distância de até 50 m das câmeras (ou até 48 m do veículo, considerando que há uma diferença de 2 m entre a parte frontal do veículo e o sistema de câmeras, conforme apresentado na Figura 18). Para que fosse possível medir a distância do pedestre em relação as câmeras, o veículo foi mantido estático durante a aquisição das imagens.

Com a finalidade de comparar o desempenho do *VeryFast* com as imagens adquiridas pelo sistema de câmeras proposto com uma câmera estéreo comercial, os dois pares de câmera estéreo foram sincronizados com a *Bumblebee* XB3 da Point Grey para realizar a aquisição de imagens limitada a 7 *fps*. Essa limitação foi considerada para que fosse possível rotular todas as imagens adquiridas e posteriormente realizar a avaliação. Como resultado, foi obtido um *dataset* com 400 *frames* por câmera (quatro câmeras monoculares para o sistema proposto, e um par de câmeras *Bumblebee*). Em seguida, todas as imagens esquerdas de cada câmera estéreo foram rotuladas manualmente através do aplicativo *Training Image Labeler* do MATLAB, que permite rotular os pedestres nas imagens de uma maneira mais simples.

Para a câmera *Bumblebee*, todos os pedestres contidos nas imagens foram rotulados. Porém, devido ao sistema estéreo proposto operar em diferentes regiões de distâncias e não haver uma integração entre as câmeras, esse processo de rotulação foi realizado de maneira diferente. Conforme foi explicado na Seção 5.1, cada câmera estéreo é responsável por operar em um intervalo de distância à frente do veículo. Por essa razão, foram rotulados para cada imagem somente os pedestres que estão no intervalo correspondente à região

de operação de cada par de câmeras. Isto é, para o sistema estéreo com *baseline* menor responsável por realizar a detecção de pedestres que estão em uma região de 5 a 20 metros à frente das câmeras, somente os pedestres que estão dentro dessa região foram rotulados. E em casos de pedestres mais distantes que 20 *m*, a rotulação foi realizada apenas para o sistema estéreo com *baseline* maior. A Figura 31 mostra um exemplo de rotulação em *frames* equivalentes obtidos pelas duas câmeras do sistema proposto. Para a câmera com distância focal de 4,5 *mm* (imagem esquerda), somente os pedestres que estão em um intervalo de 20 *m* à frente das câmeras foram rotulados, enquanto que os outros pedestres mais distantes foram rotulados nas imagens obtidas pela câmera com distância focal de 8,5 *mm* (imagem direita).

Figura 31: Pedestres rotulados em *frames* equivalentes obtidos pelas câmeras com distâncias focais de 4,5 *mm* (esquerda) e 8,5 *mm* (direita).

Fonte: Elaborado pela autora e publicado em Nakamura, Horita e Grassi Jr. (2017b).

5.4 Considerações finais

Este capítulo apresentou uma configuração de câmeras formado por dois sistemas de câmera estéreo com diferentes distâncias focais e *baselines*, a qual permitiu adquirir imagens de pedestres que estão distantes do veículo em maiores escalas. Para analisar a eficácia do sistema proposto, os pares estéreos apresentados no presente trabalho foram sincronizados com uma câmera comercial (*Bumblebee XB3*) para realizar a aquisição de um novo conjunto de dados com pedestres caminhando em uma região de até 50 metros à frente das câmeras.

6 RESULTADOS

Para evitar uma colisão, o pedestre deve ser detectado a uma distância mínima do veículo. Conforme apresentado na Seção 4, essa distância varia de acordo com o tempo de resposta do sistema de detecção de pedestres, o tempo de atuação do sistema de freio, e a dinâmica do veículo durante a frenagem (e.g., velocidade inicial, peso e tamanho do veículo). Considerando que no Brasil o limite de velocidade em uma via arterial é de 60 km/h , os pedestres devem ser detectados a pelo menos 32 m de distância à frente do veículo (ou 34 m à frente das câmeras). Porém, em alguns países o limite de velocidade em vias urbanas pode atingir 80 km/h (e.g, Coreia do Sul e Canadá), e para evitar uma colisão nesta velocidade, a mínima distância que os pedestres devem ser detectados aumenta para aproximadamente 48 m (Figura 19).

6.1 Metodologia de avaliação

Com a finalidade de avaliar quantitativamente a configuração de câmeras proposta, foram utilizados os parâmetros de *recall* e *precision*. O *recall* consiste na quantidade de pedestres contidos no *dataset* que o algoritmo conseguiu detectar corretamente, enquanto que a taxa de precisão representa a quantidade de detecções realizadas corretamente considerando todas as detecções (Figura 32).

Figura 32: *Precision e Recall*.

Fonte: Elaborado pela autora.

Após executar o detector de pedestres (*VeryFast*) no conjunto de dados coletado, a quantidade de detecções verdadeiro positivo (VP) e falso positivo (FP) foram computados em razão do limiar do coeficiente de *Jaccard*, enquanto que a quantidade de detecções

falso negativos (FN) é dado pelo número de pedestres rotulados que não foram detectados. O coeficiente de *Jaccard* é definido pela razão entre a área de intersecção e a área de união entre os *bounding box* da detecção e o *bounding box* do *ground truth*, dado por:

$$jc = \frac{D \cap GT}{D \cup GT}$$

Dessa forma, a detecção é definida como VP se o coeficiente de *jaccard* for superior ao limiar de 0,5 (valor considerado neste trabalho), caso contrário, a detecção será FP. Finalmente, os valores de *recall* e *precision* podem ser computados pelas seguintes equações:

$$recall = \frac{VP}{VP + FN}$$

e

$$precision = \frac{VP}{VP + FP}.$$

6.2 Experimentos e avaliação

A partir da necessidade do pedestre ser detectado a uma distância mínima do veículo, a avaliação dos algoritmos *VeryFast* (com e sem *stixels*) foi realizada dividindo o conjunto de dados coletado em duas partes.

A primeira parte consistiu em avaliar os algoritmos em situações que os pedestres estão caminhando em uma região de até 34 metros à frente das câmeras. Então, foram utilizados os primeiros 261 *frames* do conjunto de dados, totalizando 1084 anotações de pedestres por sistema estéreo. Para as situações em que os pedestres estão a uma distância de 34 a 50 metros à frente das câmeras, os 139 *frames* restantes foram utilizados para avaliar o desempenho dos algoritmos, totalizando 887 anotações de pedestres.

6.2.1 Parte 1

Devido ao sistema proposto ter sido executado de forma independente, a avaliação foi realizada separadamente para cada par de câmeras. Dessa forma, o valor absoluto de *recall* foi calculado a partir dos resultados obtidos por cada câmera estéreo (4,5 *mm* e 8,5 *mm*). A Tabela 3 apresenta os resultados da primeira parte dos experimentos (testes em imagens monocular e estéreo). Como pode ser observado, em situações de pedestres em uma região de até 34 metros à frente das câmeras, o sistema proposto possibilita uma melhora de pelo menos 38% no desempenho absoluto de *recall* sobre os resultados obtidos pela *Bumblebee*. E, quando é utilizado informações de profundidade (*stixels*), a melhora é maior que 45%. Esse progresso no desempenho já era esperado, uma vez que os pedestres distantes das câmeras são capturados em maiores escalas pelo sistema estéreo com lentes de 8,5 *mm*, como pode ser observado na Figura 33(a). Em relação à taxa de precisão, não foi possível obter o valor absoluto do mesmo, pois os algoritmos de detecção de pedestres

Tabela 3: Resultados das detecções - pedestres até 34 metros distância das câmeras.

	Sistema	f	VP	FP	FN	Recall	Precision
Mono	Bumblebee	3,8	547	107	537	50,46%	83,64%
	Proposto	4,5	363	273	59	86,02%	57,07%
		8,5	600	160	62	90,63%	78,95%
		Total	963	-	121	88,84%	-
Estéreo	Bumblebee	3,8	457	1	627	42,16%	99,78%
	Proposto	4,5	346	224	76	81,99%	60,70%
		8,5	602	137	60	90,94%	81,46%
		Total	948	-	136	87,45%	-

Fonte: Elaborado pela autora.

foram executados e avaliados independentemente para cada par de câmeras do sistema proposto (lentes com 4,5 mm e 8,5 mm).

É possível notar através da Tabela 3 que a taxa de precisão obtida por cada câmera estéreo proposta é inferior em comparação à *Bumblebee*. O motivo desse baixo desempenho são as detecções fora da região definida para cada par de câmeras, as quais não estão necessariamente incorretas, mas são consideradas como FPs devido a forma como os *ground truth* foram gerados para cada par de câmeras. Por exemplo, na imagem central da Figura 33(b) é visível que todos os pedestres foram detectados corretamente, porém, devido esta imagem corresponder ao sistema estéreo responsável pela detecção de pedestres que estão em uma região de até 20 metros à frente das câmeras, somente os dois pedestres que estão dentro dessa região foram rotulados e considerados como VPs. A mesma situação pode ser verificada na imagem direita da mesma figura, a qual corresponde aos pedestres que estão mais distantes que 20 metros. Dessa forma, apesar de haver uma redução na taxa de precisão em relação à *Bumblebee*, esse caso não é considerado como um problema, dado que a eficácia do sistema proposto não é prejudicado.

Além das detecções incorretas causadas pela região de operação de cada câmera, há outros casos menos frequentes de falsas detecções que são causados por objetos com aparência similar aos pedestres (imagem direita da Figura 33(a)), e detecção de apenas parte dos pedestres (Figura 33(c)). As falsas detecções devido à detecção de partes dos pedestres ocorrem quando o coeficiente de *jaccard* computado não atinge o valor limiar de 0,5 para ser considerado como VP, gerando um FP e um FN.

Nas imagens obtidas pela *Bumblebee*, a principal causa para as detecções FNs foram os pedestres em menores escalas, ou seja, aqueles que estão a uma longa distância à frente do veículo (imagens esquerda das Figuras 33(a)-(d)). Por exemplo, ao considerar o menor pedestre contido no *dataset* coletado (1,5 m de altura), é esperado que o detector comece a apresentar maior dificuldade a partir de 20 metros de distância, como pode ser observado

na Figura 21. Por outro lado, nas imagens obtidas pelo sistema proposto, as detecções FNs normalmente ocorriam em casos de razão de aspecto atípico (imagens esquerda e direita da Figura 33(d)), e em situações onde parte do corpo dos pedestres estavam sob efeito de oclusão (imagens esquerda e direita da Figura 33(e)).

Através do uso das câmeras com lentes de $8,5\text{ mm}$, foi possível reduzir a quantidade absoluta de detecções FNs em ambos os testes (monocular e estéreo) por um fator de pelo menos 4,44 (Tabela 3), o que supostamente significa que o sistema proposto superou o problema com pequenas escalas. Porém, as câmeras com menor distância focal ($4,5\text{ mm}$) ainda são necessárias para detectar os pedestres que estão mais próximos do veículo (Figura 33(f)).

Figura 33: Exemplos de detecção de pedestres que estão até 34 metros de distância das câmeras nas imagens esquerda de cada sistema estéreo (da esquerda para a direita, *Bumblebee*, $4,5\text{ mm}$ e $8,5\text{ mm}$). *Bounding boxes* (retângulos) na cor azul representam os *ground truth*, enquanto que na cor verde indicam detecções VPs, e na cor vermelha os FPs.

(a) Detecção FP causado por aparência similar ao de pedestre (imagem direita).

(b) Situações de detecções FPs causados por detecção fora da região de operação definida para cada par de câmeras (imagens central e direita).

(c) Exemplo de FP causado por detecção fora da região de operação definida para cada par de câmeras (imagem central), e FP/FN causado por detecção de parte dos pedestres, não correspondendo ao critério de *jaccard* (imagem direita).

(d) Detecções FNs causados por razão de aspecto atípica (imagens esquerda e direita).

(e) Detecções FNs causados por oclusão (imagem direita).

(f) Casos onde as câmeras com menor distância focal são necessárias.

Fonte: Elaborado pela autora e publicado em Nakamura, Horita e Grassi Jr. (2017b).

6.2.2 Parte 2

Como a segunda parte do experimento consistiu em avaliar o desempenho dos detectores somente em situações onde os pedestres estão a um intervalo de 34 a 50 metros à frente das câmeras, a câmera estéreo com lentes de 4,5 mm não foi considerada nesta avaliação. Dessa forma, a Tabela 4 apresenta apenas os resultados dos testes realizados com

Tabela 4: Resultados das detecções - pedestres em uma região de 34 m a 50 m à frente das câmeras.

	Sistema	f	VP	FP	FN	Recall	Precision
Mono	Bumblebee	3,8	26	0	861	2,93%	100,00%
	Proposto	8,5	727	183	160	81,93%	79,85%
Estéreo	Bumblebee	3,8	17	0	870	1,92%	100,00%
	Proposto	8,5	725	61	162	81,78%	92,23%

Fonte: Elaborado pela autora.

as imagens obtidas pela câmera *Bumblebee* e pelo par de câmeras com lentes de 8,5 mm. Como esperado, devido a pequena escala dos pedestres, a avaliação do detector realizada com as imagens obtidas pela câmera *Bumblebee* resultou em um desempenho muito ruim (2,93% de *recall* nos testes realizados com imagens monocular). Em compensação, a avaliação realizada com as imagens obtidas pelo sistema proposto resultou em uma taxa de *recall* de 81,93%. A Figura 34 apresenta um exemplo de pedestres a aproximadamente 50 m à frente das câmeras sendo totalmente detectados nas imagens obtidas pelo sistema proposto, e completamente falho na *Bumblebee*.

Figura 34: Exemplo do resultado da detecção em situação de pedestres a uma distância de aproximadamente 50 m à frente das câmeras. Esquerda: *Bumblebee*. Direita: sistema proposto.

Fonte: Elaborado pela autora.

Em relação aos números de FPs, com exceção das detecções fora da região definida para cada câmera, as principais causas das falsas detecções são as mesmas da Parte 1. Dessa forma, considerando que nesta segunda parte não há situação de um pedestre ser detectado por ambas as câmeras do sistema proposto (4,5 mm e 8,5 mm), percebe-se que a principal causa da baixa taxa de precisão nos testes realizados na Parte 1 foram as detecções fora da região de operação. Já a quantidade de FNs foram gerados quando em um determinado *frame*, algum pedestre deixava de ser detectado por motivo de oclusão, razão de aspecto, ou até mesmo mudança de contraste e iluminação. Esse problema com FNs poderia ser solucionado com a implementação de um algoritmo de rastreamento, pois, é possível notar que o sistema consegue detectar o pedestre na imagem, mas é somente em um *frame* que o detector falha e depois retorna a detectar (Figura 35).

Os testes foram realizados em uma máquina com Intel Core i7-4960X (16GB RAM) e NVIDIA 780i Ti, que permitiu executar os testes monocular e estéreo a uma taxa de 37 e 83 Hz, respectivamente. Apesar do uso de informações de profundidade normalmente ter resultado em uma pequena queda na taxa de *recall*, ele acelerou consideravelmente o tempo de execução por um fator de 2,24 e reduziu a quantidade de detecções FPs em

Figura 35: Situações onde o detector falhou em uma determinada sequência de *frames*.

(a) Sequência de *frames* - 11, 12 e 13

(b) Sequência de *frames* - 22, 23 e 24.

Fonte: Elaborado pela autora.

pelo menos 14%. Essas melhorias estão associadas ao fato do *stixels* delimitar a região de interesse (ROI) apenas em regiões onde é considerado a existência de obstáculos, reduzindo a quantidade de dados a serem processados e evitando regiões onde não há pedestres (Figura 36). No entanto, deve ser considerado que a configuração proposta consiste de dois sistemas estéreo, o que reduziria a taxa de execução do detector de pedestres em pelo menos metade de sua velocidade. Porém, essa velocidade ainda é o suficiente para realizar a execução em tempo real.

Figura 36: Resultado das detecções de pedestres com *stixels*.

Fonte: Elaborado pela autora.

6.3 Considerações finais

Este capítulo apresentou os resultados obtidos pela abordagem proposta. Através da avaliação e comparação do sistema proposto com a câmera comercial *Bumblebee*, foi possível observar que a configuração de câmeras proposta possibilitou uma melhora na taxa de *recall* de pelo menos 38% em situações de pedestres a uma distância de até 34 metros à frente das câmeras. E em situações de pedestres em um intervalo de 34 a 50 metros, a melhora foi de pelo menos 79%. Apesar dos testes com informações de profundidade terem resultado em uma taxa de *recall* um pouco menor do que os testes realizados utilizando apenas imagens monoculares, o uso de imagens estéreo traz grandes vantagens em relação a melhora na taxa de precisão e, principalmente, na rápida execução. Através dos resultados obtidos, conclui-se que a configuração de câmeras proposta possibilitou superar a principal causa das falsas detecções em aplicações automotivas, que consiste em detectar pedestres que estão mais distantes do veículo, devido suas pequenas escalas nas imagens.

7 CONCLUSÃO

Devido ao seu potencial em salvar vidas, extensas pesquisas são realizadas no ramo de detecção de pedestres desde o final da década de 90. Em comparação aos humanos, a aplicação destes sistemas em veículos traz grandes vantagens em relação ao tempo de reação. Conforme foi apresentado na Figura 22, ao utilizar um sistema de detecção de pedestres para assistir o motorista, a distância percorrida durante o tempo de reação é 45 vezes menor em comparação ao tempo de reação de um humano.

No entanto, apesar da relevante evolução na qualidade das detecções, ainda não é possível obter simultaneamente um bom desempenho com a execução em tempo real, devido à complexidade dessa tarefa e às limitações atuais de *hardware* e *software*. Como pode ser visto nos *benchmarks*, os métodos que utilizam arquiteturas profundas de redes neurais possuem os melhores desempenhos, mas os métodos baseados em janelas deslizantes são muito mais rápidos. Assim, para saber se essa evolução é o suficiente para aplicações automobilísticas e tornar o trânsito mais seguro, foi realizado um estudo para analisar a aplicabilidade dos métodos estado-da-arte e verificar a qual distância os pedestres deveriam ser detectados para que um veículo conseguisse frear a tempo.

Através da análise, ao considerar uma câmera estéreo comercial (*Bumblebee XB3*), verificou-se que ainda não se pode confiar totalmente nos métodos visuais de detecção de pedestres para assistir o motorista. Pois, os métodos deveriam ser capazes de detectar pedestres a uma distância de pelo menos 34 metros à frente das câmeras, e o melhor método avaliado no *benchmark* da Caltech deixa de detectar 77% dos pedestres nesta situação devido às suas pequenas escalas na imagem. Porém, para pedestres em maiores escalas, os detectores já conseguem atingir taxas de erro praticamente nulas. Ademais, a diferença de desempenho entre os métodos que utilizam arquiteturas profundas de redes neurais e aqueles baseados em janelas deslizantes é pequena, enquanto que a diferença no tempo de execução é muito significativa. Dessa forma, nesse trabalho buscou-se uma maneira de permitir o uso de detectores visuais de pedestres para assistir o motorista através da restrição de operação dos métodos apenas para as escalas onde os detectores conseguem atingir um bom desempenho, e ao mesmo tempo com rápida execução. Isto foi feito através da proposta de uma configuração de câmeras que utiliza dois pares estéreo com diferentes distâncias focais e *baselines*.

A fim de analisar a eficácia do sistema proposto e compará-lo com uma câmera comercial, um novo conjunto de dados foi coletado utilizando o sistema de câmeras proposto sincronizado com a câmera estéreo *Bumblebee XB3* da Point Grey. Por fim, a avaliação foi realizada executando o detector *VeryFast* nas imagens obtidas por ambos sistemas estéreos. Como consequência, o sistema proposto possibilitou alcançar resultados promissores e

consideravelmente melhores se comparado com as imagens adquiridas pela *Bumblebee*, especialmente em relação aos pedestres que estão mais distantes do veículo.

Apesar da taxa de *recall* ter sido um pouco melhor nos testes realizados com imagens monoculares, o uso de informações estéreo permitiu a execução muito mais rápida e com maior taxa de precisão. Dessa forma, neste trabalho, o uso de configurações de câmeras estéreo se mostrou mais vantajoso do que câmeras monoculares para detecção de pedestres.

7.1 Trabalhos futuros

Devido a operação independente de cada câmera estéreo da configuração proposta, não foi possível calcular a taxa absoluta da precisão. Dessa forma, para trabalhos futuros seria interessante realizar a integração das câmeras do sistema proposto, e com isso adotar um método para calcular a taxa absoluta de precisão. Além de relizar a integração das câmeras, pode-se realizar também a implementação da configuração proposta utilizando câmeras de baixo custo, e com isso, verificar a contribuição da combinação de câmeras com diferentes *baselines* e distâncias focais ao utilizar câmeras comercialmente mais acessíveis.

Até o momento foram sugeridas formas de melhorar o desempenho da detecção de pedestres a partir da configuração de hardware, porém há também recursos de software para este fim. Por exemplo, a implementação de algoritmos de rastreamento visual poderia reduzir a quantidade de detecções FNs devido a situações de oclusão, razão de aspecto atípico, variações de contraste ou iluminação, etc. No entanto, os pedestres podem mudar o percurso repentinamente, o que torna o rastreamento em uma ferramenta não totalmente confiável. Portanto, um método de estimação da intenção do pedestre poderia ser utilizado como um complemento do rastreamento. Assim, com a capacidade de detectar os pedestres e estimar suas posições futuras em relação às suas possíveis intenções, ações preventivas em casos de riscos de colisão poderão ser tomadas.

REFERÊNCIAS

- BADINO, H.; FRANKE, U.; PFEIFFER, D. The Stixel World - A Compact Medium Level Representation of the 3D-World. p. 1–10, 2009. ISSN 0302-9743.
- BENENSON, R.; MATHIAS, M.; TIMOFTE, R.; Van Gool, L. Fast stixel computation for fast pedestrian detection. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 7585 LNCS, n. PART 3, p. 11–20, 2012. ISSN 03029743.
- _____. Pedestrian detection at 100 frames per second. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, p. 2903–2910, 2012. ISSN 10636919.
- BENENSON, R.; OMRAN, M.; HOSANG, J.; SCHIELE, B. Ten years of pedestrian detection, what have we learned? In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.: s.n.], 2015. v. 8926, p. 613–627. ISBN 9783319161808. ISSN 16113349.
- BENENSON, R.; TIMOFTE, R.; Van Gool, L. Stixels estimation without depth map computation. **Proceedings of the IEEE International Conference on Computer Vision**, p. 2010–2017, 2011.
- CAI, Z.; FAN, Q.; FERIS, R. S.; VASCONCELOS, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. p. 1–16, 2016.
- CHEN, G.; DING, Y.; XIAO, J.; HAN, T. X. Detection evolution with multi-order contextual co-occurrence. In: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2013. p. 1798–1805. ISBN 978-0-7695-4989-7. ISSN 10636919.
- Companhia de Engenharia de Tráfego. Acidentes de Trânsito: Relatório anual 2016. p. 70, 2017. Disponível em: <<http://www.cetsp.com.br/media/562061/relatorioanualacidentestransito-2016.pdf>>.
- COSTEA, A. D.; NEDEVSCI, S. Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 2393–2400. ISBN 9781479951178. ISSN 10636919.
- DALAL, N.; TRIGGS, B. Histograms of Oriented Gradients for Human Detection. In: **2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. [S.l.: IEEE, 2005. v. 1, p. 886–893. ISBN 0-7695-2372-2.
- Digital Photography Review. **Making (some) sense out of sensor sizes: Digital Photography Review**. 2002. Disponível em: <<https://www.dpreview.com/articles/8095816568/sensorsizes>>.

DOLLÁR, P.; APPEL, R.; BELONGIE, S.; PERONA, P. Fast feature pyramids for object detection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 36, n. 8, p. 1532–1545, 2014. ISSN 01628828.

DOLLÁR, P.; BELONGIE, S.; PERONA, P. The Fastest Pedestrian Detector in the West. In: **Proceedings of the British Machine Vision Conference 2010**. [S.l.]: British Machine Vision Association, 2010. p. 68.1–68.11. ISBN 1-901725-40-5.

DOLLÁR, P.; TU, Z.; PERONA, P.; BELONGIE, S. Integral Channel Features. In: **Proceedings of the British Machine Vision Conference 2009**. [S.l.]: British Machine Vision Association, 2009. p. 91.1–91.11. ISBN 1-901725-39-1.

DOLLÁR, P.; TU, Z.; TAO, H.; BELONGIE, S. Feature mining for image classification. In: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2007. ISBN 1424411807. ISSN 10636919.

DOLLÁR, P.; WOJEK, C.; SCHIELE, B.; PERONA, P. Pedestrian detection: A benchmark. **2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009**, p. 304–311, 2009. ISSN 1063-6919.

_____. 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, apr 2012. ISSN 0162-8828.

DU, X.; EL-KHAMY, M.; LEE, J.; DAVIS, L. S. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. **Computing Research Repository**, oct 2016.

ENZWEILER, M.; GAVRILA, D. M. Monocular pedestrian detection: Survey and experiments. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 31, n. 12, p. 2179–2195, 2009. ISSN 01628828.

ESS, A.; LEIBE, B.; Van Gool, L. **Depth and Appearance for Mobile Scene Analysis**. 2007. 1–8 p.

FRAUNDORFER, F.; SCARAMUZZA, D. Visual Odometry Part II: Matching, Robustness, Optimization, and Applications. **IEEE Robotics & Automation Magazine**, p. 80–92, dec 2012. ISSN 1070-9932.

GANDHI, T.; TRIVEDI, M. M. Pedestrian protection systems: Issues, survey, and challenges. **IEEE Transactions on Intelligent Transportation Systems**, v. 8, n. 3, p. 413–430, sep 2007. ISSN 15249050.

GARDINALLI, G. J. **Comparação do Desempenho de Frenagem Simulada x Experimental de um Veículo de Passeio com Freios Hidráulicos e ABS**. 2005. 96 p. Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo, 2005. Disponível em: <http://automotiva-poliusp.org.br/wp-content/uploads/2013/02/gardinalli{_}gerald0.>

GEIGER, A.; LENZ, P.; URTASUN, R. Are we ready for autonomous driving? the KITTI vision benchmark suite. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, p. 3354–3361, 2012. ISSN 10636919.

GERÓNIMO, D.; LÓPEZ, A. M. **Vision-based Pedestrian Protection Systems for Intelligent Vehicles**. New York, NY: Springer New York, 2014. 114 p. (SpringerBriefs in Computer Science). ISBN 978-1-4614-7986-4.

GERÓNIMO, D.; LÓPEZ, A. M.; SAPPA, A. D.; GRAF, T. Survey of Pedestrian Detection for Advanced Driver Assistance Systems. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 7, p. 1239–1258, jul 2010. ISSN 0162-8828.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, p. 580–587, 2014. ISSN 10636919.

GONZÁLEZ, A.; FANG, Z.; SOCARRAS, Y.; SERRAT, J.; VÁZQUEZ, D.; XU, J.; LÓPEZ, A. Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison. **Sensors**, v. 16, n. 6, p. 820, jun 2016. ISSN 1424-8220.

HIRSCHMÜLLER, H.; SCHARSTEIN, D. Evaluation of Cost Functions for Stereo Matching. **2007 IEEE Conference on Computer Vision and Pattern Recognition**, p. 1–8, 2007. ISSN 1063-6919. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4270273>>.

JOHANSSON, G.; RUMAR, K. Drivers' brake reaction times. **Human factors**, v. 13, n. 1, p. 23–27, 1971. ISSN 0018-7208.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. **Advances In Neural Information Processing Systems**, p. 1–9, 2012. ISSN 10495258.

KUBOTA, S.; NAKANO, T.; OKAMOTO, Y. A Global Optimization Algorithm for Real-Time On-Board Stereo Obstacle Detection Systems. **Proc. IEEE Intelligent Vehicles Symposium**, p. 7–12, 2007. ISSN 1931-0587.

LABAYRADE, R.; AUBERT, D.; TAREL, J.-P. Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation. **Intelligent Vehicle Symposium, 2002. IEEE**, v. 2, p. 646–651, 2002. ISSN 19449224.

LEIBE, B.; SEEMANN, E.; SCHIELE, B. Pedestrian detection in crowded scenes. In: **IEEE. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. [S.l.], 2005. v. 1, p. 878–885.

LI, J.; LIANG, X.; SHEN, S.; XU, T.; YAN, S. Scale-aware Fast R-CNN for Pedestrian Detection. **arXiv preprint**, p. 1–9, 2015.

MARIN, J.; VAZQUEZ, D.; LOPEZ, A. M.; AMORES, J.; LEIBE, B. Random forests of local experts for pedestrian detection. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2013. p. 2592–2599. ISBN 9781479928392. ISSN 1550-5499.

NAKAMURA, A. T. M.; HORITA, L. R. T.; GRASSI Jr., V. A Aplicabilidade dos Métodos Visuais de Detecção de Pedestres para Veículos Inteligentes. In: **XIII Simpósio Brasileiro de Automação Inteligente**. Porto Alegre: [s.n.], 2017. p. 1616–1622.

_____. A stereo cameras setup for pedestrian detection enhancement. In: **2017 IEEE International Symposium on Robotics and Intelligent Sensors (IEEE-IRIS2017)**. Ottawa, Canada: [s.n.], 2017.

NAM, W.; DOLLÁR, P.; HAN, J. H. Local Decorrelation For Improved Detection. **Nips**, p. 1–9, 2014. ISSN 10495258.

OKUTOMI, M.; KANADE, T. A Multiple-Baseline Stereo. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 15, n. 4, p. 353–363, 1993. ISSN 01628828.

PAISITKRIANGKRAI, S.; SHEN, C.; HENGEL, A. V. D. Efficient Pedestrian Detection by Directly Optimizing the Partial Area under the ROC Curve. **2013 IEEE International Conference on Computer Vision**, p. 1057–1064, 2013. ISSN 1550-5499.

PAPAGEORGIOU, C.; POGGIO, T. Trainable system for object detection. **International Journal of Computer Vision**, v. 38, n. 1, p. 15–33, 2000. ISSN 09205691.

PEPIKBO, B.; BENENSON, R.; RITSCHER, T.; SCHIELE, B. What is holding back convnets for detection? **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 9358, p. 517–528, 2015. ISSN 16113349.

PREMEBIDA, C.; LUDWIG, O.; NUNES, U. Exploiting LIDAR-based features on pedestrian detection in urban scenarios. In: **2009 12th International IEEE Conference on Intelligent Transportation Systems**. [S.l.]: IEEE, 2009. p. 1–6. ISBN 978-1-4244-5519-5.

REN, J.; CHEN, X.; LIU, J.; SUN, W.; PANG, J.; YAN, Q.; TAI, Y.-W.; XU, L. Accurate single stage detector using recurrent rolling convolution. In: **Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017.

ROSEIRO, L. A. R. **Detecção de obstáculos usando fusão de dados de percepção 3D e radar em veículos automotivos**. 2017. 85 p. Tese (Doutorado) — Universidade de São Paulo, 2017.

SHINZATO, P. Y.; SANTOS, T. C. dos; ROSEIRO, L. A.; RIDEL, D. A.; MASSERA, C. M.; ALENCAR, F.; BATISTA, M. P.; HATA, A. Y.; OSÓRIO, F. S.; WOLF, D. F. Carina dataset: An emerging-country urban scenario benchmark for road detection systems. In: **2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)**. [S.l.: s.n.], 2016. p. 41–46.

SZARVAS, M.; SAKAI, U.; OGATA, J. Real-time Pedestrian Detection Using LIDAR and Convolutional Neural Networks. In: **2006 IEEE Intelligent Vehicles Symposium**. [S.l.]: IEEE, 2006. p. 213–218. ISBN 4-901122-86-X.

SZELISKI, R. **Computer Vision: Algorithms and Applications**. [S.l.]: Springer Science & Business Media, 2010. 812 p. ISBN 1848829353.

TRUCCO, E.; VERRI, A. **Introductory Techniques for 3-D Computer Vision**. [S.l.]: Prentice Hall, 1998. 343 p. ISBN 0132611082.

VIOLA, P.; JONES, M. J. Robust real-time face detection. **International Journal of Computer Vision**, v. 57, n. 2, p. 137–154, 2004. ISSN 09205691.

WALK, S.; MAJER, N.; SCHINDLER, K.; SCHIELE, B. New features and insights for pedestrian detection. In: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2010. p. 1030–1037. ISBN 9781424469840. ISSN 10636919.

WANG, X.; HAN, T. X.; YAN, S. An HOG-LBP human detector with partial occlusion handling. **Computer Vision, 2009 IEEE 12th International Conference on**, n. Iccv, p. 32–39, 2009. ISSN 1550-5499.

WOJEK, C.; SCHIELE, B. A performance evaluation of single and multi-feature people detection. In: **Pattern Recognition**. [S.l.]: Springer Berlin Heidelberg, 2008. v. 5096, p. 82–91. ISBN 978-3-540-69320-8.

WOJEK, C.; WALK, S.; SCHIELE, B. Multi-cue onboard pedestrian detection. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.]: IEEE, 2009. p. 794–801. ISBN 978-1-4244-3992-8. ISSN 1063-6919.

WONG, J. Y. **Theory of Ground Vehicles**. 4. ed. Hoboken, New Jersey: John Wiley & Sons, 2008. 592 p. ISBN 9780471354611.

ZHANG, S.; BENENSON, R.; OMRAN, M.; HOSANG, J.; SCHIELE, B. How Far are We from Solving Pedestrian Detection? **Computer Vision and Pattern Recognition (CVPR)**, 2016.

ZHANG, S.; BENENSON, R.; SCHIELE, B. Filtered Channel Features for Pedestrian Detection. **Computer Vision and Pattern Recognition (CVPR)**, p. 1751–1760, 2015. ISSN 10636919.

_____. CityPersons: A Diverse Dataset for Pedestrian Detection. feb 2017.

ZHANG, Z. A flexible new technique for camera calibration. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 11, p. 1330–1334, Nov 2000. ISSN 0162-8828.

Apêndices

APÊNDICE A – PUBLICAÇÃO

NAKAMURA, A. T. M.; HORITA, L. R. T.; GRASSI Jr., V. Aplicabilidade dos Métodos Visuais de Detecção de Pedestres para Veículos Inteligentes. In: **XIII Simpósio Brasileiro de Automação Inteligente (SBAI)**. Porto Alegre: 2017. p. 1616-1622.

NAKAMURA, A. T. M.; HORITA, L. R. T.; GRASSI Jr., V. A Stereo Cameras Setup for Pedestrian Detection Enhancement. In: **IEEE 5th International Symposium on Robotics and Intelligent Sensors (IEEE-IRIS2017)**. Ottawa, Canada: 2017.