AVALIAÇÃO DE ACURÁCIA DE REDES NEURAIS CONVOLUCIONAIS NA DETECÇÃO DE VEÍCULOS E PEDESTRES NO TRÂNSITO BRASILEIRO

Pedro Augusto Pinho Ferraz* Bernardo Augusto Godinho de Oliveira* Thiago Melo Machado-Coelho† Álvaro Henrique de Araújo Rungue* Willian Antônio dos Santos* Flávia Magalhães Freitas Ferreira* Carlos Augusto Paiva da Silva Martins*

*Pontifícia Universidade Católica de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica
Av. Itaú 525, 30535-012, Belo Horizonte, MG, Brasil

†Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

Email: pferraz@sga.pucminas.br, bernardo.godinho@sga.pucminas.br, thmmcoelho@ufmg.br, alvaro.rungue@sga.pucminas.br, willian.santos.838460@sga.pucminas.br, flaviamagfreitas@pucminas.br, capsm@pucminas.br

Abstract— One of the most cricital aspects in the context of autonomous vehicles development is how capable the system is to correcly identify obstacles, whether vehicles, pedestrians, cyclists or any others. With the rise of Deep Learning and Convolutional Neural Networks, detection solutions based on image processing using only conventional cameras have been shown to be economically viable and efficient. In addition, there are many available datasets with thousands of images for training and validation of these nets. However, the datasets are usually made in the USA, Germany or other countries that don't necessarily portray transit scenarios from emerging countries, like Brazil. This paper evaluates the accuracy in the detection of vehicles and pedestrians of state-of-the-art algorithms, comparing the results of KITTI dataset with a new dataset, proposed in this work, from the city of Belo Horizonte. The networks behaved in a lower way when subjected to the images of Belo Horizonte, reaching an accuracy of 12.91% lower in vehicles and 20.12% lower for pedestrians,if compared to the results from KITTI test dataset. These results indicate that these regional differences may have a significant impact on how well the algorithm will detect obstacles.

Keywords— Convolutional Neural Network, Deep Learning, Vechile Detection, Obstacle Detection, Object Recognition.

Resumo— Um dos pontos mais críticos no contexto de veículos autônomos é a capacidade de identificar corretamente os obstáculos, sejam veículos, pedestres, ciclistas ou outros. Com o advento do *Deep Learning* e das Redes Neurais Convolucionais, soluções de detecção utilizando apenas câmeras convencionais têm se mostrado economicamente viáveis e eficientes. Além disso, existem *datasets* específicos para veículos autônomos, com milhares de imagens disponíveis para treinamento e validação. Entretanto, estes *datasets* geralmente são produzidos nos EUA, Alemanha ou outros países que não retratam, necessariamente, cenários de trânsito de países emergentes, como o Brasil. Este trabalho realiza uma avaliação dos resultados de acurácia na detecção de veículos e pedestres de algoritmos estado da arte, comparando os resultados utilizando o *dataset* KITTI com um novo *dataset* da cidade de Belo Horizonte, proposto neste trabalho. As redes se comportaram de maneira inferior quando submetidas às imagens de Belo Horizonte, chegando a ter uma acurácia 12.91% menor em veículos e 20.12% menor para pedestres, se comparados com a acurácia na base de testes KITTI. Essa variação indica que diferenças regionais podem ter impactos significativos em quão bem o algoritmo irá detectar os obstáculos.

Palavras-chave— Redes Neurais Convolucionais, Aprendizado Profundo, Detecção de Veículos, Detecção de Obstáculos, Reconhecimento de Objetos.

1 Introdução

Nas últimas décadas, os veículos têm se tornado cada vez mais eletrônicos do que mecânicos, devido aos constantes avanços da microeletrônica, automação e computação embarcada. Como consequência, uma grande variedade de sensores, distribuídos ao longo do veículo, alimentam algoritmos que têm como objetivo aprimorar a condução e tornar o automóvel mais robusto e seguro. Sistemas como ABS, piloto automático, direção elétrica e assistente de subida em rampa são algumas dessas aplicações.

A área de visão computacional aplicada à segurança e à assistência do motorista não foi diferente e tem motivado pesquisas há mais de 20 anos, como os trabalhos da *Subaru Research Center* (Saneyoshi, 1994)(Saneyoshi, 1996). Com os recentes investimentos no desenvolvimento de tecnologias para veículos autônomos, somados ao aumento considerável de capacidade computacional atual, a área teve avanços significativos, principalmente com o desenvolvimento de técnicas para identificação e segmentação de obstáculos. Vários veículos que possuem capacidade de direção autônoma ou assistência avançada baseiam-se em sensores de custo elevado, tais como LiDAR *(Light Detection and Ranging)*(Wolcott e Eustice, 2017), radares (Cho e Tseng, 2013), infravermelho (Suard et al., 2006)(Kwak et al., 2017)

e outros (Chellappa et al., 2004), além de câmeras tradicionais. Entretanto, com o advento do *Deep Learning* e das Redes Neurais Convolucionais (CNN - *Convolutional Neural Network*) (LeCun et al., 1998) (Krizhevsky et al., 2012), soluções utilizando apenas câmeras convencionais têm se mostrado cada vez mais economicamente viáveis e eficientes. Dessa forma, existe um constante esforço em pesquisa e desenvolvimento para aprimorar as Redes Neurais Convolucionais, tanto no quesito desempenho computacional quanto em acurácia nas tarefas de detecção de objetos.

Existem, atualmente, diversas empresas e universidades coletando dados e gerando *datasets* para treinamento dessas redes, tais como KITTI (Geiger et al., 2013) e Citiscapes (Cordts et al., 2016). Todavia, vários deles são obtidos em outros países, tais como EUA, Alemanha, Japão, dentre outros, não necessariamente refletindo os cenários e particularidades de trânsito de outras localidades, principalmente em países emergentes como o Brasil.

O objetivo deste trabalho é avaliar a aplicabilidade de Redes Neurais Convolucionais treinadas em outros *datasets* aplicadas no reconhecimento de obstáculos no cenário de trânsito brasileiro. Como estudo de caso, será utilizado o *dataset* KITTI, originalmente obtido do trânsito na Alemanha, para treinar as redes a serem aplicadas nas imagens na cidade de Belo Horizonte, Minas Gerais.

Como primeira contribuição, é apresentado o *Dataset* BH, composto de imagens com anotações da cidade de Belo Horizonte. Outra contribuição é o estudo inicial sobre o comportamento de algumas Redes Neurais Convolucionais consideradas estado da arte, treinadas com o conjunto de imagens fornecido pelo KITTI, quando aplicadas ao cenário de trânsito brasileiro. Esse estudo avalia a acurácia e a capacidade de generalização destas redes na detecção de obstáculos nas imagens obtidas pelos autores deste trabalho.

Este trabalho foi estruturado da seguinte maneira: na Seção 2 é feita uma revisão bibliográfica e apresentados os principais trabalhos relacionados; na Seção 3 é apresentado o *Dataset* BH, proposto para testes das redes treinadas com o *dataset* KITTI; já na Seção 4, os experimentos são descritos; na Seção 5 são apresentados e discutidos os resultados, comparando a acurácia das redes originais com os testes no novo *dataset*; por fim, as conclusões e considerações são discutidas na Seção 6.

2 Trabalhos Relacionados

Nos últimos anos, o *Deep Learning* consolidou-se como o estado da arte em vários problemas de computação, tais como processamento de linguagem natural, reconhecimento de voz, tradução, predição e também visão computacional (Krizhevsky et al., 2012). Dentre as diversas arquiteturas de *Deep Learning*, podemos citar as Redes Neurais Convolucionais, foco deste trabalho.

A competição ILSVRC *(ImageNet Large Scale Visual Recognition Competition)* (Deng et al., 2009)(Russakovsky et al., 2015), realizada anualmente, é uma das mais importantes na área de visão computacional e tem como objetivo medir o estado da arte dos algoritmos em tarefas de detecção de objeto e classificação de imagem através de um *dataset* contendo mais de 14 milhões de imagens, divididos entre 10 mil categorias. Até 2011, todos os algoritmos que competiam eram baseados em características, ou seja, os autores escolhiam as características mais relevantes e as respectivas técnicas de extração, agrupavam os dados gerados e executavam um classificador, criando uma arquitetura de alta complexidade e com taxas de erro na ordem de 25% (Yuanqing Lin, 2011). Porém, em 2012, um trabalho publicado na mesma competição mudou os paradigmas na área de visão computacional. O modelo proposto por Alex Krizhevsky, chamado AlexNet (Krizhevsky et al., 2012) foi a primeira CNN prática utilizando GPU e conquistou o primeiro lugar na competição com 16,4% de taxa de erro, cerca de 10% menor que qualquer outro participante daquela edição. A partir desse trabalho, todos os vencedores do ILSVRC e aplicações em imagem passaram a utilizar redes neurais convolucionais, em suas diversas configurações.

No contexto de detecção de obstáculos para a navegação de veículos autônomos, existem algumas particularidades que demandam um desenvolvimento mais direcionado. Existem vários algoritmos utilizados para detecção de objetos de maneira genérica, como Overfeat (Sermanet et al., 2013), Faster-RCNN (Ren et al., 2015), SPP (He et al., 2014), YOLO (Redmon e Farhadi, 2016)(Redmon et al., 2016), cada um com uma abordagem diferente. Entretanto, nenhum deles representa o estado da arte na detecção de obstáculos para a navegação de veículos autônomos, de acordo com o *benchmark* KITTI (Geiger et al., 2012), principalmente pelo fato de não conseguirem detectar nem localizar, de forma adequada, objetos muito pequenos ou com grandes oclusões; além disso, podem estar obstruídos por outros objetos, estar refletindo ou sendo refletido, dentre outras possibilidades. Dessa forma, nem sempre o estado da arte da competição ILSVRC será também o estado da arte em *datasets* como este. Uma dessas diferenças é a grande variação de escala dos objetos, ilustrada pela Fig. 1 (Cai et al., 2016), que foi extraída do KITTI, um dos *datasets* mais utilizados (Geiger et al., 2012).

Proposto através da parceria do Karlsruhe Institute of Technology com o Toyota Technological Institute, o KITTI (Geiger et al., 2012) foi concebido como um *dataset* e *benchmark* de vi-

Figura 1: Vários veículos em diferentes escalas em uma única imagem.

são computacional na área de veículos autônomos, a fim de impulsionar o desenvolvimento na área. Gravado na cidade de Karlsruhe, na Alemanha, o KITTI é amplamente utilizado, possuindo diversas métricas para avaliação de algoritmos, tais como visão estéreo, odometria visual, detecção de objeto e rastreamento de objetos.

Para a categoria de detecção de objetos, o *dataset* disponibiliza dois conjuntos de dados: um para treinamento e validação, contendo 7481 imagens e outro para teste, utilizado somente para *benchmark* oficial, que possui 7518 imagens. Dentro da base de treinamento/validação, foram anotados 51865 objetos, divididos em várias classes mas, para fins de acurácia, considera apenas três: veículos (28742 objetos), pedestres (4487 objetos) e ciclistas (1627 objetos). Para cada uma dessas classes, existem níveis de dificuldade na detecção, segmentados em *easy*, *moderate* e *hard*, de acordo com seu tamanho e quão obstruído está na imagem. A métrica de acurácia, implementada no KITTI, avalia cada uma dessas classes em cada um dos níveis. Entretanto, para classificar os algoritmos no *ranking Object Evaluation KITTI*, utiliza os resultados do nível *moderate*. Para atuar nessa questão, redes como Multi-scale CNN (MSCNN)(Cai et al., 2016), Subcategory-aware CNN (SubCNN) (Xiang et al., 2016), Subcategory-aware CNN (RRC) (Ren et al., 2017) e várias outras têm sido constantemente desenvolvidas. Assim, o *ranking* de *Object Evaluation* KITTI é frequentemente atualizado, ditando o novo estado da arte.

Uma área menos explorada é a aplicabilidade e análise desses algoritmos em cenários reais, em condições adversas, como a análise empírica de (Huval et al., 2015). Sendo assim, o presente trabalho propõe uma avaliação dos algoritmos estado da arte em detecção de obstáculos no KITTI, quando aplicados em imagens de trânsito brasileiro, analisando, assim, sua acurácia e capacidade de generalização neste novo conjunto de dados. Como estudo de caso, serão utilizadas imagens de trânsito capturadas na cidade de Belo Horizonte.

3 Dataset BH para Avaliação de CNNs

Com a finalidade de avaliar a acurácia de detecção e capacidade de generalização das CNNs treinadas previamente com *datasets* de veículos autônomos, foi criada uma base de imagens, ainda em estado inicial, chamada *Dataset* BH[1], no âmbito deste trabalho. O *dataset* ainda está em processo de captura de imagens e, assim que forem obtidos um volume maior de imagens anotadas, será disponibilizado publicamente. Esta seção apresentará o detalhamento de geração e processamento do *Dataset* BH.

3.1 Aquisição dos Dados

Para obter as amostras de imagem do trânsito em Belo Horizonte, uma webcam convencional, modelo Logitech C270, foi instalada no topo do veículo de testes. A câmera foi conectada a um computador, responsável pela coleta e armazenamento das imagens, obtidas a partir de cenários diversos, variando as condições de trânsito. Por questões de privacidade, todas as placas de veículos e faces de pedestres foram desfocadas. As imagens foram capturadas no formato PNG, utilizando resolução de 1280x720 pixels. O *setup* de aquisição de dados é ilustrado pela Fig. 2.

Figura 2: Camera montada no veículo de testes.

O percurso gravado foi de aproximadamente 30 km, passando por áreas distintas de Belo Horizonte, considerando locais movimentados, avenidas e vias principais, áreas residenciais, rodovias e áreas de caminhada. Do total de imagens capturadas, foram selecionadas 255 para passarem pelo processo de anotação e serem utilizadas nos testes. Cenários diferentes, grande quantidade de obstáculos e variação de tamanho e obstrução foram os critérios utilizados para escolher as imagens, similares aos utilizados pelo KITTI. Após selecionadas, as imagens foram redimensionadas e recortadas para possuírem a mesma resolução das imagens disponibilizadas pelo KITTI, 1242x375 pixels. As proporções das imagens do *Dataset* BH foram mantidas, para não interferirem no processo de detecção.

3.2 Processo de Anotação de Objetos

Seguindo os mesmos critérios de preparação e anotação dos dados do *dataset* KITTI, as imagens fo-

[1] https://github.com/ferrazpedro1/dataset_bh

ram rotuladas em três categorias: Veículos, Pedestres e Ciclistas. As anotações foram feitas através do software ALPS Labeling Tool (ALPS Labeling Tool, n.d.). Cada objeto presente na imagem é classificado e é desenhada uma caixa delimitadora *(bounding box)* que, posteriormente, será comparada com a saída das redes de detecção, assim como em (de Oliveira et al., 2018). Portanto, será possível saber se a CNN localizou e classificou corretamente os objetos. O total de objetos catalogados foi de 1295, sendo 997 veículos, 281 pedestres e 17 ciclistas. Apesar de ser uma quantidade inferior de imagens e objetos anotados, se comparados com o KITTI, as proporções de objetos por imagem são maiores: 3.91 veículos por imagem, no *Dataset* BH, contra 3.84 veículos por imagem no KITTI; 1.10 pedestres por imagem, contra 0.60 no KITTI.

Algumas dessas imagens estão ilustradas com suas devidas caixas delimitadoras na Fig. 3.

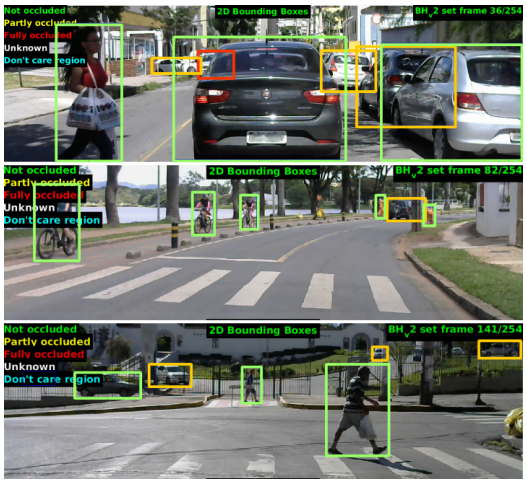

Figura 3: Exemplo de imagens do *Dataset* BH.

3.3 Aplicação

O *Dataset* BH gerado será utilizado como base de testes para avaliar a acurácia e capacidade de generalização de CNNs treinadas em *datasets* próprios da área de veículos autônomos. O *dataset* KITTI, que será utilizado para fins de comparação, fornece 7481 imagens para treinamento, todas com anotações dos objetos, e 7518 imagens para teste, utilizadas pela ferramenta de *benchmark*[2].

Apesar das diferenças regionais, um bom algoritmo deve ser capaz de generalizar as características dos objetos em imagens nunca apresentadas para a rede, identificando veículos, mesmo que de outros fabricantes e modelos, e também pedestres. A Fig. 4 mostra um exemplo de imagem do KITTI e outro do *Dataset* BH. Ao avaliar a acurácia das CNNs ao identificar objetos nas imagens do *Dataset* BH, tendo as CNNs sido treinadas com as imagens do KITTI, as redes devem ser capazes de reconhecer os mesmos tipos de obstáculos. O quão bem elas farão este reconhecimento será discutido nas próximas seções.

[2] `http://www.cvlibs.net/datasets/kitti/eval_object.php`

Figura 4: Exemplo de imagem do *Dataset* KITTI (cima) e similar do *Dataset* BH (baixo).

4 Experimentos

Para avaliar a acurácia das CNNs na detecção de obstáculos do novo *Dataset* BH, foram selecionadas algumas redes consideradas estado da arte em detecção de veículos e pedestres pelo *Object Evaluation Ranking*. Os critérios de escolha das redes foram baseadas na sua acurácia publicada no *ranking* oficial KITTI, disponibilidade de código-fonte e também considerando limitações do nosso hardware disponível. Existem outras CNNs que poderiam ser utilizadas nos testes, entretanto, ou ela extrapolaria a memória disponível em nossa GPU ou seu treinamento demoraria muitos dias. Desta forma, as redes selecionadas foram a Faster-RCNN (Ren et al., 2015), a MSCNN (Cai et al., 2016), a SubCNN (Xiang et al., 2016) e a RRC (Ren et al., 2017). A Tabela 1 mostra a colocação dessas redes convolucionais no *ranking* de Detecção de Objeto KITTI, em março de 2018.

Tabela 1: Colocação CNNs Ranking KITTI

Colocação Ranking KITTI Detecção de Objeto			
Modelo	**Carros**	**Pedestres**	**Ciclistas**
RRC	5°	7°	4°
SubCNN	27°	17°	13°
MSCNN	30°	12°	5°
Faster-RCNN	61°	30°	25°

Todas as redes foram treinadas utilizando as 7481 imagens da base de treino KITTI, e a acurácia utilizada para efeitos de comparação é extraída do resultado oficial publicado no KITTI *Object Evaluation Ranking*, que valida a acurácia através de 7518 imagens, cujas anotações não são publicamente disponíveis. As redes MSCNN e RRC disponibilizam os pesos pré-treinados das redes submetidas para o *ranking* oficial, de forma que não foi necessário realizar nenhum processo de treinamento local. Já a Faster-RCNN e a SubCNN não

disponibilizaram os pesos, por isso foram treinadas em hardware próprio, mas utilizando as mesmas parametrizações dos trabalhos originais, para efeitos de comparação justos. Utilizando as mesmas redes treinadas no mesmo conjunto de imagens, é feita a análise comparativa de detecção de obstáculos nas 255 imagens do *Dataset* BH, que possuem mesma dimensão das imagens do KITTI, não sendo necessária nenhuma adaptação em nenhuma das CNNs. Tanto o processo de treinamento e detecção das redes é feito utilizando o *framework* de *deep learning* Caffe (Jia et al., 2014). Os detalhes de hardware e software utilizados são descritos pela Tabela 2.

Tabela 2: Configuração de hardware e software utilizados nos testes.

Hardware e software utilizados	
CPU	Intel Core i3 6100 3.7GHz
RAM	16GB DDR4
GPU	Nvidia Geforce 1070 8GB
OS	Ubuntu 16.04 64bits
OpenCV	3.3.0
CUDA	8.0
CUDNN	7.1.4

Para avaliar as redes, serão utilizadas três métricas: (1) Acurácia medida em cada um dos níveis (*easy*, *moderate* e *hard*), para as detecções KITTI e *Dataset* BH; (2) Variação de Acurácia: a diferença entre o valor obtido nas detecções KITTI e das obtidas no *Dataset* BH; (3) Variação Média de Acurácia: a média aritmética da variação de acurácia nos níveis *easy*, *moderate* e *hard*. Os resultados serão apresentados e discutidos na próxima seção.

5 Resultados

Os resultados apresentados comparam a acurácia de detecção de obstáculos pelas redes MSCNN, Faster-RCNN, SubCNN e RRC nas imagens de teste do KITTI com as imagens do *Dataset* BH. Foram executadas as detecções apenas para as categorias Veículos e Pedestres, uma vez que não foram capturadas amostras suficientes de ciclistas pelo *Dataset* BH para realizar uma comparação efetiva. As métricas de acurácia de detecção, em ambos os casos, são as disponibilizadas pelo KITTI. Os resultados originais das CNNs, denotadas pelo sufixo *-KITTI*, estão publicamente disponíveis no *ranking* oficial KITTI.

5.1 Detecção de Veículos

As acurácias na detecção de veículos pelas redes são apresentados na Tabela 3.

Os resultados indicam que as redes apresentaram acurácia inferior quando testadas no *Dataset* BH, se comparadas com o *dataset* KITTI em quase todos os níveis. Esta redução foi mais significativa na Faster-RCNN, chegando a acurácia ser 12.94% menor no nível *moderate*. A Faster-RCNN também teve acurácia 12.61% menor no nível *easy*, e 9.17% no nível *hard*, tendo sido a rede com a maior variação negativa dentre as testadas. Ao contrário do esperado, a SubCNN teve um ganho no nível *easy*, de 0.53%, enquanto nos outros níveis há redução, sendo a maior delas no nível *moderate*, com 9.37%. A MSCNN teve redução nos níveis *easy* and *moderate*, mas registrou um ganho considerável no nível *hard*, de 2.33%. Já a RRC, que é a rede melhor ranqueada no *Object Evaluation Ranking* dentre as avalidadas, apresentou uma variação negativa, mas muito pequena no nível *easy*, sendo de apenas 0.07%, mantendo a acurácia neste nível próximo dos 90%. A variação de acurácia no nível *moderate* foi a menor das redes testadas, sendo de apenas -1.46%, mantendo a RRC como a melhor ranqueada, dentre as testadas, uma vez que o KITTI classifica as redes pelo nível *moderate*.

Tabela 3: Resultados Acurácia Veículos

Acurácia Detecção Carros (%)			
Rede	**Easy**	**Moderate**	**Hard**
MSCNN 8s-768 - KITTI	90.46	88.83	74.76
MSCNN-8s-768 - BH	89.09	83.83	77.09
Variação Acurácia:	**-1.37**	**-5.00**	**+2.33**
Faster-RCNN - KITTI	87.90	79.11	70.19
Faster-RCNN - BH	75.29	66.17	61.02
Variação Acurácia:	**-12.61**	**-12.94**	**-9.17**
SubCNN - KITTI	90.75	88.86	79.24
SubCNN - BH	91.28	79.49	75.61
Variação Acurácia:	**+0.53**	**-9.37**	**-3.63**
RRC - KITTI	90.61	90.22	87.44
RRC - BH	90.54	88.76	80.84
Variação Acurácia:	**-0.07**	**-1.46**	**-6.60**

Quando comparadas pela variação de acurácia média, a RRC foi a que teve melhor resultado (-2.71%), enquanto as Faster-RCNN foi a com pior variação, de -11.57%. A MSCNN e a SubCNN tiveram variações médias próximas, -2.90% e -4.16%, respectivamente. Esses resultados sugerem que as redes MSCNN e RRC possuem uma melhor capacidade de generalização, se comparadas com a Faster-RCNN.

5.2 Detecção de Pedestres

A Tabela 4 mostra os resultados de acurácia para a classe Pedestres. Não foi possível avaliar a RRC nessa classe, pois os autores não disponibilizaram os pesos pretreinados e o *hardware* utilizado nesses testes não é suficiente para seu treinamento.

A CNN que teve a maior redução em acurácia ao detectar pedestres foi a SubCNN, chegando a 20.12% no nível *easy*, registrando resultados menores do que a própria Faster-RCNN. A MSCNN teve a menor redução dentre as testadas; entretanto, chegou a ser quase 10% menor, quando analisado no nível *easy*. Todas as redes tiveram

Tabela 4: Resultados Acurácia Pedestres

Acurácia Detecção Pedestres (%)			
Rede	**Easy**	**Moderate**	**Hard**
MSCNN 8s-768 - KITTI	83.70	73.62	68.28
MSCNN-8s-768 - BH	74.38	68.28	65.97
Variação Acurácia:	**-9.32**	**-5.34**	**-2.31**
Faster-RCNN - KITTI	78.35	65.91	61.19
Faster-RCNN - BH	62.20	57.72	56.65
Variação Acurácia:	**-16.15**	**-8.19**	**-4.54**
SubCNN - KITTI	83.17	71.34	66.36
SubCNN - BH	63.05	56.37	54.98
Variação Acurácia:	**-20.12**	**-14.97**	**-11.38**

reduções mais severas para pedestres do que para veículos.

Avaliando a média de redução da acurácia, na MSCNN ela foi 5.66% menor e na Faster-RCNN ficou em 9.63%, enquanto que na SubCNN foi 15.49%, quando comparadas com as redes testadas com o KITTI. Considerando a capacidade de generalização, a MSCNN se mostrou mais robusta, seguida da Faster-RCNN e, por fim, a SubCNN. Um fator que pode ter contribuído para uma robustez maior da MSCNN, se comparada com as outras redes testadas, é o fato do autor ter optado por treinar duas redes distintas: uma para detecção apenas de veículos e outra para detecção de pedestres e ciclistas.

6 Conclusões

O estudo inicial apresentado neste trabalho indica uma possível redução na acurácia na detecção de veículos e pedestres em imagens de localidades diferentes das utilizadas no treinamento, considerando o estudo específico de redes treinadas no KITTI e submetidas ao *Dataset* BH. Esta questão é importante pois está diretamente relacionada com o funcionamento de veículos autônomos submetidos a ambientes muito diferentes dos apresentados durante o treinamento. Dentre as redes analisadas, a SubCNN foi a que teve o resultado mais deteriorado, seguido da Faster-RCNN, enquanto a RRC e a MSCNN tiveram uma queda menor na acurácia de detecção, mostrando uma melhor capacidade de generalização.

O *Dataset* BH provê exemplos de imagens do trânsito brasileiro e pode servir de parâmetro para analisar algoritmos treinados em outros conjuntos de dados. Abordagens futuras incluem realizar novos testes utilizando uma versão expandida e colaborativa do *Dataset* BH, incluindo mais cidades brasileiras e cenários de trânsito diversos.

Outra possibilidade de trabalho futuro é a de executar um processo de retreino nestas redes, utilizando a transferência de conhecimento do KITTI com estas novas imagens do *Dataset* BH para avaliar o impacto na acurácia final.

Agradecimentos

Os autores agradecem à CAPES por suportarem esta pesquisa através do seu programa de bolsas acadêmicas.

Referências

ALPS Labeling Tool (n.d.). Alps labeling tool.

Cai, Z., Fan, Q., Feris, R. S. e Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection, CoRR **abs/1607.07155**.

Chellappa, R., Qian, G. e Zheng, Q. (2004). Vehicle detection and tracking using acoustic and video sensors, Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, Vol. 3, IEEE, pp. iii–793.

Cho, H.-J. e Tseng, M.-T. (2013). A support vector machine approach to cmos-based radar signal processing for vehicle classification and speed estimation, Mathematical and Computer Modelling **58**(1-2): 438–448.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. e Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

de Oliveira, B. A. G., Ferreira, F. M. F. e da Silva Martins, C. A. P. (2018). Fast and lightweight object detection network: Detection and recognition on resource constrained devices, IEEE Access **6**: 8714–8724.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. e Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database, CVPR09.

Geiger, A., Lenz, P., Stiller, C. e Urtasun, R. (2013). Vision meets robotics: The kitti dataset, International Journal of Robotics Research (IJRR) .

Geiger, A., Lenz, P. e Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite, Conference on Computer Vision and Pattern Recognition (CVPR).

He, K., Zhang, X., Ren, S. e Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition, european conference on computer vision, Springer, pp. 346–361.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R. et al. (2015). An empirical evaluation of deep learning on highway driving, arXiv preprint arXiv:1504.01716 .

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. e Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093 .

Krizhevsky, A., Sutskever, I. e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks, in F. Pereira, C. J. C. Burges, L. Bottou e K. Q. Weinberger (eds), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp. 1097–1105.

Kwak, J.-Y., Ko, B. C. e Nam, J. Y. (2017). Pedestrian tracking using online boosted random ferns learning in far-infrared imagery for safe driving at night, IEEE Transactions on Intelligent Transportation Systems **18**(1): 69–81.

LeCun, Y., Bottou, L., Bengio, Y. e Haffner, P. (1998). Gradient-based learning applied to document recognition, Proceedings of the IEEE **86**(11): 2278–2324.

Redmon, J., Divvala, S., Girshick, R. e Farhadi, A. (2016). You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Redmon, J. e Farhadi, A. (2016). Yolo9000: Better, faster, stronger, arXiv preprint arXiv:1612.08242 .

Ren, J., Chen, X., Liu, J., Sun, W., Pang, J., Yan, Q., Tai, Y.-W. e Xu, L. (2017). Accurate single stage detector using recurrent rolling convolution, Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, pp. 752–760.

Ren, S., He, K., Girshick, R. B. e Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks, CoRR **abs/1506.01497**.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. e Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) **115**(3): 211–252.

Saneyoshi, K. (1994). 3-d image recognition system by means of stereoscopy combined with ordinary image processing, Intelligent Vehicles' 94 Symposium, Proceedings of the, IEEE, pp. 13–18.

Saneyoshi, K. (1996). Drive assist system using stereo image recognition, Intelligent Vehicles Symposium, 1996., Proceedings of the 1996 IEEE.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. e LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks, CoRR **abs/1312.6229**.

Suard, F., Rakotomamonjy, A., Bensrhair, A. e Broggi, A. (2006). Pedestrian detection using infrared images and histograms of oriented gradients, 2006 IEEE Intelligent Vehicles Symposium, IEEE, pp. 206–212.

Wolcott, R. W. e Eustice, R. M. (2017). Robust lidar localization using multiresolution gaussian mixture maps for autonomous driving, The International Journal of Robotics Research **36**(3): 292–319.

Xiang, Y., Choi, W., Lin, Y. e Savarese, S. (2016). Subcategory-aware convolutional neural networks for object proposals and detection, CoRR **abs/1604.04693**.

Yuanqing Lin, Fengjun Lv, S. Z. M. Y. T. C. K. Y. L. C. Z. L. M.-H. T. X. Z. T. H. T. Z. (2011). Large-scale image classification: Fast feature extraction and svm training, CVPR .