COMPARAÇÃO ENTRE MÉTODOS DE DETECÇÃO DE OBJETOS POR PONTOS-CHAVE E REDES NEURAIS

Pedro Henrique Silva Domingues[1], Douglas De Rizzo Meneghetti[2], Plinio Thomaz Aquino Junior[3]
Departamento de Ciência da Computação, FEI
12pedro07@gmail.com, douglasrizzo@fei.edu.br, plinio.aquino@fei.edu.br

Resumo: O projeto envolve estudos comparativos entre os sensores Kinect v1 e v2 e seus comportamentos quando relacionados com diferentes algoritmos de detecção de objetos, sendo estes: SIFT, RootSIFT e redes convolucionais. Na pesquisa também podem ser encontrados estudos sobre as diferenças entre estes três algoritmos, apontando suas vantagens e desvantagens de cada um e estabelecendo vínculos de como estes fatores podem influenciar no desempenho do robô no qual a pesquisa foi aplicada.

1. Introdução

A detecção de objetos é um dos mais importantes atributos da robótica móvel de assistência doméstica, uma vez que uma das principais atividades destes robôs é o transporte de objetos e, para isto, é necessário identificar posição e distância destes no ambiente que está inserido. Por esse motivo, os algoritmos de visão computacional têm sido foco de pesquisa dos últimos anos, o que resultou em avanços consideráveis na área de *machine learning*, que ainda assim apresenta alguma nebulosidade sobre seus resultados e ainda podendo ser comparados a algoritmos mais tradicionais como SIFT [1] e RootSIFT [2], sob algumas condições.

Este projeto visa melhorar o desempenho da área de visão computacional do robô de pesquisa e competição HERA, da equipe RoboFEI, mais especificamente sua funcionalidade de detecção de objetos. Para isso, é necessário manter controle das limitações provenientes do ambiente competitivo durante a pesquisa, por exemplo, a inviabilidade na obtenção de amostras muito grandes de imagens dos objetos ou o processamento muito lento, uma vez que o algoritmo funcionar é aplicado a um vídeo em tempo real no qual objetos podem aparecer por apenas alguns *frames*. O robô possui um conjunto de sensores que são utilizados para navegação, manipulação, reconhecimento de objetos e interação com humanos. Para o reconhecimento de objetivos, utilizam-se câmeras de vídeo digital e sensores Kinect. Dado esse contexto de sensores do robô, adicionalmente, esta pesquisa determinou qual é a melhor combinação entre sensor Kinect (v1 ou v2) com cada um dos algoritmos (SIFT, RootSIFT e MobileNet [3] com *single shot detector* [4]), com foco em produzir os resultados mais proveitosos para um robô móvel.

2. Metodologia

Para realização das comparações envolvendo os três algoritmos, 100 *frames* foram extraídos de vídeos gravados com os sensores Kinect v1 e v2 (50 *frames* de cada sensor) contendo um conjunto de 13 classes de objetos diferentes. A tabela I mostra a proporção na qual cada classe é representada nas imagens. Estas imagens foram posteriormente utilizadas para realizar os testes e estudos comparativos, para os quais foi montado um gabarito, no qual consta a informação relativa à presença ou ausência de cada objeto em cada uma destas imagens.

Tabela I – Presença de objeto na amostra de 100 imagens.

Objetos	Presença nas imagens
cereal	49%
chocolate_milk	51%
heineken	40%
iron_man	44%
medicine	52%
milk_bottle	58%
milk_box	53%
monster	53%
purple_juice	42%
red_juice	41%
shampoo	47%
tea_box	59%
yellow_juice	47%

Para os algoritmos SIFT e RootSIFT, foram tiradas fotos de todas as faces de cada um dos objetos, as quais foram tratadas para ficarem em escala de cinza e em escalas (bases) de 5%, 10%, 25%, 50%, 75% e 100%.

Para a rede convolucional, foram acumuladas 166 imagens, as quais podem conter qualquer quantidade dos objetos selecionados, variando de zero (objeto ausente na imagem), até o máximo de duas cópias deste na mesma foto. Todos os objetos foram marcados nas 166 imagens para serem utilizados no treinamento da rede. Com isto, foram feitos testes nas 100 amostras citadas anteriormente utilizando os três algoritmos.

3. Resultados

A comparação entre os algoritmos e sensores Kinect demonstrou diferenças claras entre a aplicação de uma rede neural com poucas amostras de treinamento e os algoritmos SIFT e RootSIFT, pois, como pode ser visto na Figura 1, a velocidade de processamento (independente do sensor) é incomparavelmente menor ao utilizarmos a rede neural, porém o contraponto pode ser percebido através da Figura 2, na qual a precisão relativa (medida criada para relacionar o número de objetos encontrados na imagem pelo algoritmo, com o número real de objetos presentes) é muito maior nos outros dois algoritmos mesmo não sendo desprezível para a rede convolucional.

Com relação às comparações entre parâmetros utilizados no SIFT e RootSIFT, foram estudadas as escalas utilizadas como base (5%, 10%, 25%, 50%, 75% e 100%) e o número mínimo de pontos comuns entre a imagem de um objeto e da imagem a ser analisada, para

que o algoritmo aponte um objeto como presente em uma imagem. Em ambos os casos, a precisão relativa encontra-se sempre proporcional ao número de falsos positivos detectados, como exemplificado nas Figuras 3 e 4.

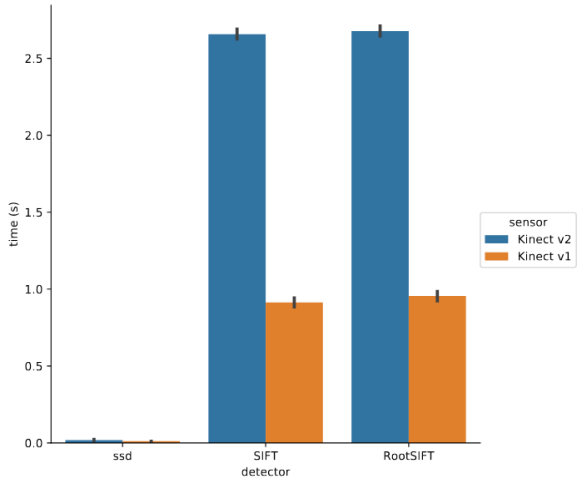

Figura 1- Tempo para uma detecção x Detector e Sensor

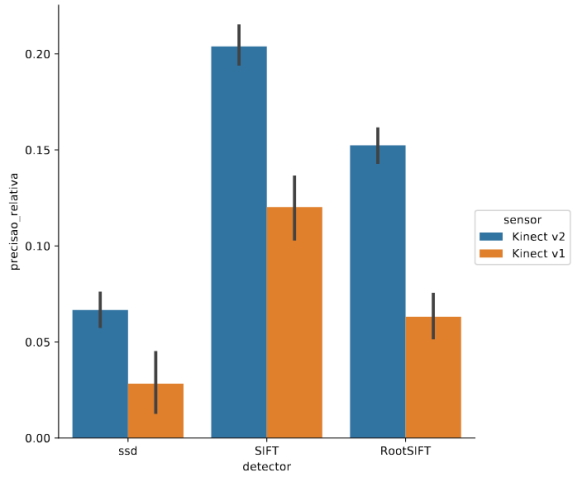

Figura 2 – Precisão Relativa x Detector e Sensor

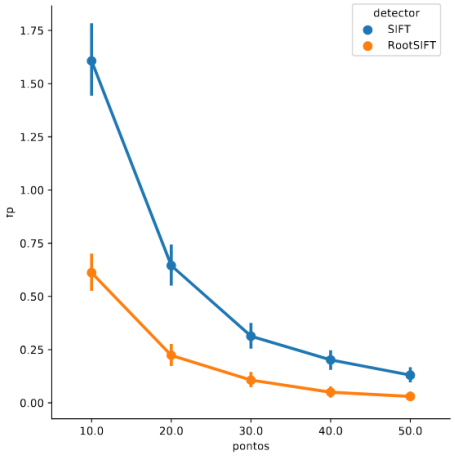

Figura 3 – Falsos positivos x número de pontos.

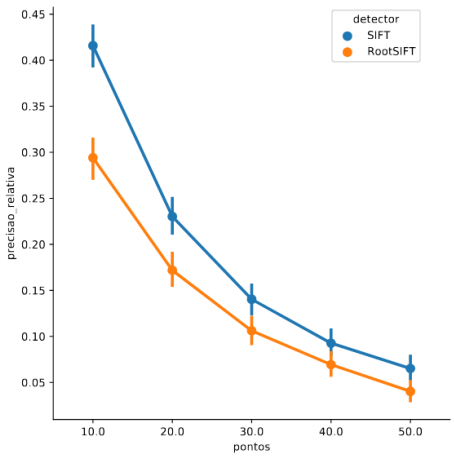

Figura 4 – Precisão relativa x número de pontos

4. Conclusões

A partir das comparações realizadas durante a pesquisa, foi decidido que por apresentar melhor velocidade de processamento, maior abertura para pesquisas e para melhoria de desempenho, a rede neural será o algoritmo implementado no robô.

A desvantagem encontrada no método escolhido é a menor precisão, problema este que pode ser resolvido por meio de pesquisas futuras e testes com variações de parâmetros como número de imagens utilizadas para treinamento, repetibilidade do cenário de fundo nestas imagens, parâmetros internos da rede, etc.

Em relação ao sensor escolhido, o Kinect v2 apresentou resultados mais precisos com relação ao número de acertos sobre os objetos presentes na imagem, independente do método de detecção utilizado e pouca variação no tempo de processamento quando utilizado com o algoritmo escolhido.

5. Referências

[1] LOWE,D.G. Distinctive image features fromscale-invariant keypoints.International journal of computer vision, Springer, v. 60, n. 2, p. 91–110, 2004. Acesso em: 20 jan. 2018.[2] ARANDJELOVIC, R.; ZISSERMAN, A. Three things everyone should know to improve object retrieval. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2012. p. 2911–2918. ISSN 1063-6919. Acesso em: 10 jul. 2018.

[3] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, *arXiv:1704.04861 [cs]*, abr. 2017.

[4] W. Liu *et al.*, “SSD: Single Shot MultiBox Detector”, *arXiv:1512.02325 [cs]*, vol. 9905, p. 21–37, 2016.

Agradecimentos

Ao Centro Universitário FEI pela realização das medidas e empréstimo de equipamentos da equipe RoboFEI.

[1] Aluno de IC do CNPq. Projeto com vigência de 07/17 a 07/18.

[2] Aluno de Doutorado do Programa de Pós-Graduação em Engenharia Elétrica do Centro Universitário da FEI.

[3] Professor orientador do projeto do Centro Universitário FEI.