Deep Learning aplicado à estimativa de profundidade em imagens estéreo multivista

Sérgio M. M. de Oliveira[a,b,d], Erikson F. de Morais[a,b], Marcella S. R. Martins[a,c]

[a] *Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil*
[b] *Departamento Acadêmico de Informática - DAINF*
[c] *Departamento Acadêmico de Eletrônica - DAELE*
[d] *Autor para correspondência: sergiomurilo@alunos.utfpr.edu.br*

Palavras-chaves: visão estéreo multivista, automação veicular, aprendizado profundo

Estimar a distância entre o ponto de observação e os objetos no ambiente é parte crucial no desenvolvimento de diversas aplicações robóticas.

Atualmente, o LiDAR (*Light Detection and Ranging*) é um dos principais sensores responsáveis pela estimativa de profundidade. Com margem de erro de 2 centímetros a 200 metros do alvo nos modelos mais avançados, fornece informações precisas quanto à profundidade dos objetos nas cenas observadas. Em contrapartida, o alto custo de produção inviabiliza a utilização em larga escala do LiDAR, pois a necessária redundância de sensores em veículos autônomos, por exemplo, os torna comercialmente inviáveis.

Imagens digitais, monoculares e binoculares (captadas a partir de duas câmeras observando a mesma a cena de pontos de vista ligeiramente distintos), apresentam indicadores visuais (*depth cues*). A aplicação de técnicas de Visão Computacional no processamento de tais imagens permite inferir a profundidade da cena observada. Ainda que possuam alta resolução, câmeras digitais são mais acessíveis e baratas em comparação ao LiDAR. Por outro lado, o custo computacional, diretamente proporcional à resolução das imagens processadas, e a baixa precisão em comparação ao LiDAR dificultam a utilização de tais técnicas em aplicações comerciais.

Observando tal cenário, o presente trabalho aplica *Deep Learning* na obtenção do mapa de profundidade da cena observada. Os experimentos atuais, realizados com imagens estéreo provenientes do *DrivingStereo*[1] *Dataset* atingiram 54 % de acurácia, medida a partir da comparação entre o mapa de profundidade real (fornecido pelo *dataset*) e o mapa gerado pelo modelo. O modelo atual consiste em uma Rede Neural Siamesa com 2 sub-redes compostas por camadas convolucionais, responsáveis por processar a imagem esquerda e direita simultaneamente. O resultado passa por uma camada de concatenação e é utilizado como semente para uma Rede GAN (*Generative Adversarial Network*), que gera o mapa de profundidade da cena observada.

[1] Yang, Guorun, et al. "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

ISSN:2526-1371 DOI:10.5281/zenodo.6090322