

Depth Estimation Using Deep Neural Networks

Muşat Bogdan-Adrian

December 2017

Problem Formulation

- Given a pair of stereo images, how to estimate the depth of the existing objects?

Disparity

- Disparity represents the difference of perspective created by the horizontal or vertical separation of two cameras
- The human brain processes disparity information from both eyes to estimate the depth of real world objects
- Disparity is inverse proportional with depth; given the baseline distance b between the cameras and the camera focal length f , the depth \hat{d} from the predicted disparity d is simply $\hat{d} = bf/d$
- Objects closer to the camera have bigger disparities, whilst objects that are farther have smaller disparities

Proposed Solution

- Given a pair of rectified ¹ stereo images as input, build a model which can accurately predict the disparity per-pixel (i.e. learn the Δ values with which every pixel from the left image is shifted from the right one on the x axis)
- Such a system can be modeled by a Convolutional Neural Network

¹Vertically aligned

Challenges - Textureless Areas

- Depth estimation is an ill-posed problem: a pixel in a textureless area from the left image can belong to multiple pixels in the right image



Figure 1: Example of objects with textureless areas

Challenges - Object Occlusion

- Parts of an object may be visible in an image, while being absent in the other one because of the difference of perspective

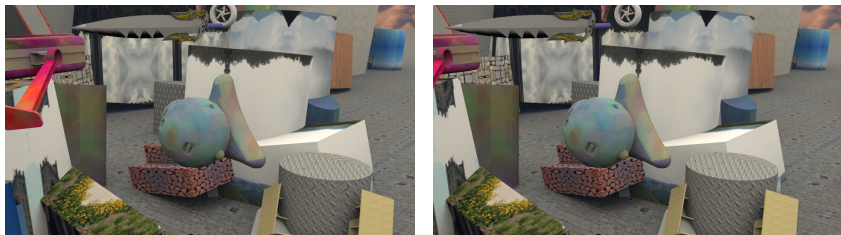


Figure 2: Occlusion in stereo images

Convolutional Neural Networks Architectures

- Supervised
 - Requires ground-truth disparity, which might be expensive to obtain
 - Based on disparity regression
 - Accurate predictions
- Unsupervised
 - Does not require any form of ground-truth
 - Based on an image reconstruction loss
 - Has artefacts around the edges of objects, caused by occlusion

DispNet² - Overview

- Image-to-image autoencoder structure

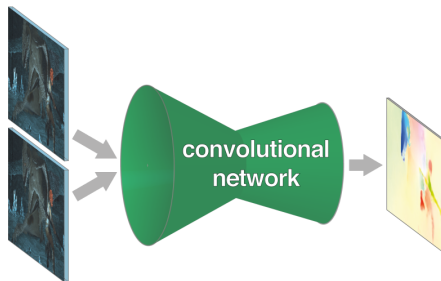


Figure 3: Hourglass structure of DispNet

²Nikolaus Mayer et al. "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation". In: *CoRR* abs/1512.02134 (2015). arXiv: 1512.02134. URL: <http://arxiv.org/abs/1512.02134>.

DispNet - Architectural Details

- Fully Convolutional Network³
- Concatenates the “upconvolution” results with the features from the “contractive” part to recover spatial information lost by downsampling
- Uses smooth L1 loss between the resulted disparity computed at different pyramid levels of the network and the ground-truth

$$|d|_{smooth} = \begin{cases} 0.5d^2, & \text{if } |d| \leq 1 \\ |d| - 0.5, & \text{otherwise} \end{cases}$$

³Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *CoRR* abs/1411.4038 (2014). arXiv: 1411.4038. URL: <http://arxiv.org/abs/1411.4038>.

DispNet - Architectural Details

- The upsampling part of the network can be constructed by using transposed convolutions⁴ or by using interpolation methods like nearest neighbor or bilinear followed by a convolutional layer to smoothen the results

⁴Matthew D. Zeiler, Graham W. Taylor, and Rob Fergus. “Adaptive Deconvolutional Networks for Mid and High Level Feature Learning”. In: *Proceedings of the 2011 International Conference on Computer Vision. ICCV '11*. Washington, DC, USA: IEEE Computer Society, 2011, pp. 2018–2025. ISBN: 978-1-4577-1101-5. DOI: [10.1109/ICCV.2011.6126474](http://dx.doi.org/10.1109/ICCV.2011.6126474). URL: <http://dx.doi.org/10.1109/ICCV.2011.6126474>.

DispNet - Results

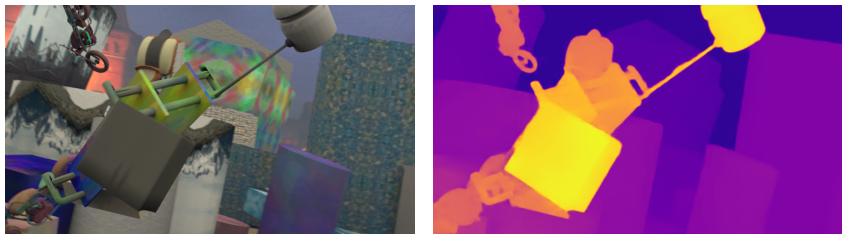


Figure 4: Supervised depth estimation

Unsupervised Monocular Network⁵ - Overview

- Similar architecture as DispNet

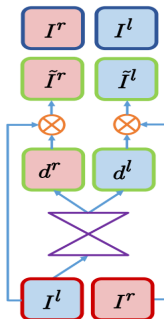


Figure 5: Unsupervised Monocular Network architecture

⁵Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised Monocular Depth Estimation with Left-Right Consistency". In: *CoRR* abs/1609.03677 (2016). arXiv: 1609.03677. URL: <http://arxiv.org/abs/1609.03677>.

Unsupervised Monocular Network - Architectural Details

- Shares the same hourglass structure as DispNet and works the same, up until a point
- Instead of using ground-truth disparity, the network tries to reconstruct the right image from the left one, and vice-versa
- The network tries to achieve the needed disparities as to warp the left and right image using some sampling method
- The sampling method chosen is bilinear sampling as used in Spatial Transformer Networks⁶, which is fully differentiable
- Produces disparity estimates at four scales

⁶Max Jaderberg et al. "Spatial Transformer Networks". In: *CoRR* abs/1506.02025 (2015). arXiv: 1506.02025. URL: <http://arxiv.org/abs/1506.02025>.

Unsupervised Monocular Network - Architectural Details

- Uses a three term loss

$$C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr}^l + C_{lr}^r)$$

- Appearance Matching Loss

$$C_{ap}^l = \frac{1}{N} \sum_{i,j} \alpha \frac{1 - SSIM(I_{ij}^l, \tilde{I}_{ij}^l)}{2} + (1 - \alpha) \|I_{ij}^l - \tilde{I}_{ij}^l\|$$

- Disparity Smoothness Loss

$$C_{ds}^l = \frac{1}{N} \sum_{i,j} |\partial_x d_{ij}^l| e^{-|\partial_x I_{ij}^l|} + |\partial_y d_{ij}^l| e^{-|\partial_y I_{ij}^l|}$$

- Left-Right Disparity Consistency Loss

$$C_{lr}^l = \frac{1}{N} \sum_{i,j} |d_{ij}^l - d_{ij+d_{ij}^l}^r|$$

Unsupervised Monocular Network - Results

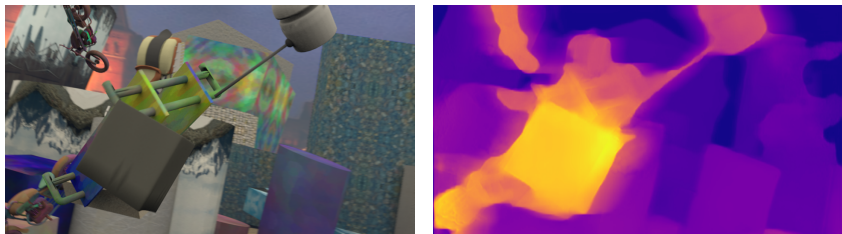


Figure 6: Unsupervised depth estimation