Varun Kumar
Summer Research Fellowship 2015
Registration Number: ENGT151
varun.invent@gmail.com

Report

Real time depth map generation from a pair of stereo images.

Aim:

- To develop a stereo vision algorithm which could generate a dense depth map in realtime.
- Which is independent of interoccular (baseline) distance.
- Which learns the deph cues from data (No hand-engineering the features).
- Robust to the major failure cases.

Stereo vision algorithms work by establishing correspondences between pixels of two images seen by the two cameras. That is, which pixel in the image taken by the left camera correspond to which pixel in the image taken by the right camera. Then the disparity is the amount by which the pixel has shifted in the second image. Using this disparity we can create the depth/disparity map.

Depth/disparity map is a representation of the relative depth of a scene seen by a camera. A depth map is generally represented by a greyscale image. Highest pixel values corresponds to objects which are closer and lowest pixel values corresponds to farthest objects.

As we are dealing with two cameras, baseline or interoccular distance is the distance between them.

Failure cases:

- Textureless surface
- Specular highlights

Textureless surface creates problem because if the surface does not have any texture then it is very difficult to match the pixels.

Specular highlights are also known as reflections. The disparity is erroneous at these surfaces because the reflection shifts and changes shape in the second image.

Related work:

- [1] suggested that depth cues can be learnt from data itself.
- Used a model based on 'Synchrony autoencoder' to learn depth /features/weights in an unsupervised way.
- They used these weights to extract depth from a stereo image pair.
- The depth map generated was sparse and failed in textureless regions.

.....

- [2] have trained a Convolutional Neural Network (CNN) with image patches to predict how well two patches match (Correspondence problem) to create a coarse depthmap.
- Coarse depthmap is refined afterwards by cross-based cost aggregation and semiglobal matching.
- Left-Right consistency check for occluded regions.

- Poor performance where specular highlight is present.

.....

- [3] have improved the performance on specular surfaces by detecting cars in the scene.
- Optimized Conditional Random Field energy function for depth estimation.
- For cars, they matched the resulting depthmap with the depthmap artificially generated by constructing the 3D CAD model of cars and adjusted the depth map by generating score based on level of agreement.
- Uses 265 sec for calculating disparity of one stereo pair. Optimization based depth estimation tends to be slow.
- Modelling of 'n' number objects that may appear is tough in realtime.

.....

- [4] Took a single image for depth map prediction.
- Trained CNN with full images along with the corresponding Depth map as Ground Truth.
- Used 2 CNNs:
    - First one predicts a coarse depthmap
    - Second one uses both the images and the coarse depthmap generated previously to predict the refined depthmap
- Fails at specular surfaces.

<u>Proposed Method</u>:

The proposed method is divided into two parts.

1. Finding the coarse discontinuous depth map using CNN.

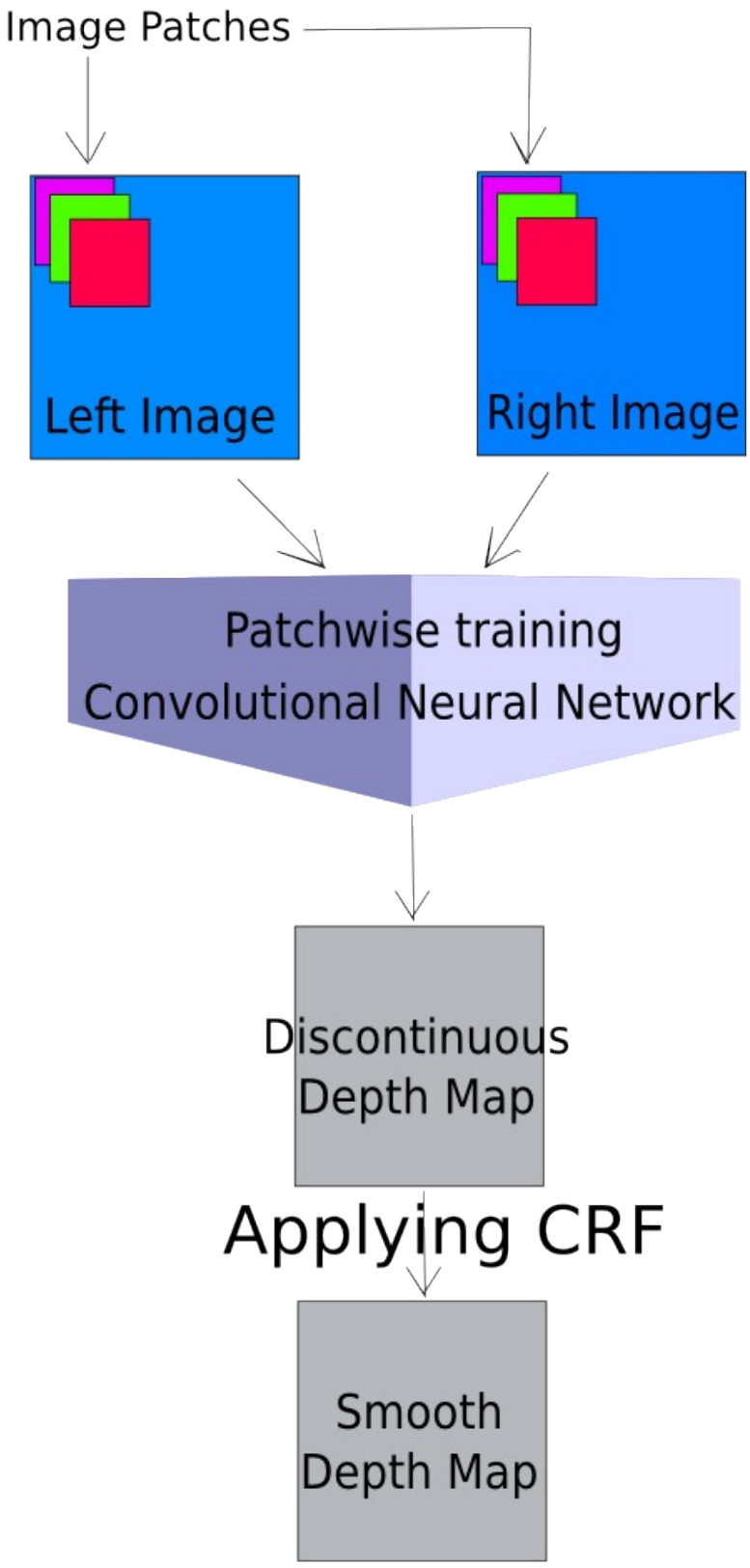
    - For each stereo image pair, let the size of images pairs be MxMx3
    - Inflate the dataset by constructing pyramids of images and ground truth (the given depth map) and then extracting the overlapping patches of size NxN from each level of these pyramids.
    - Train CNN with these overlapping patches taking 'k' kernels of size nxn and ground truth which is the original depthmap of the image as the output.
    - Then feedforward the patches from the test image pair to get the coarse discontinuous depthmap.

2. Refine the depthmap using CRF.

    - Train the CRF using the depth map calculated in the first stage and the available ground truth as the output.
    - Then refine the depthmap during the testing phase using the above trained CRF.

The proposed architecture is given on the next page.

Proposed Architecture



References:

[1] Konda, Kishore, and Roland Memisevic. "Unsupervised learning of depth and motion." arXiv preprint arXiv:1312.3429 (2013).

[2] Žbontar, Jure, and Yann LeCun. "Computing the stereo matching cost with a convolutional neural network." arXiv preprint arXiv:1409.4326 (2014).

[3] Güney, Fatma, and Andreas Geiger. "Displets: Resolving Stereo Ambiguities using Object Knowledge." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[4] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in Neural Information Processing Systems. 2014.