Disparity Estimation using Group Wise Correlation Network for Stereo imaging

by Anirudh Iyengar K N and Durga Prasad
(ENG16CS0013) (ENG16CS0028)

Guided by :
Mrs.ANUSHA.M.S
Asst.Prof, Department of CSE
Dayananda Sagar University

February 2020

Content

- Introduction
- Problem Definition
- Functional Requirement
- Module
- Low level design
- Flowchart
- Test Cases
- Result Analysis
- Conclusion

Introduction

- Disparity estimation refers to the set of techniques and algorithms aiming to obtain a representation of the spatial structure of a scene. In other terms, to obtain a measure of the distance of, ideally, each point of the scene seen.
- Correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. Group-wise correlation provides efficient representations for measuring feature similarities and will not lose too much information like full correlation.
- Stereo imaging is that Two (or more) cameras (images), with the depth information extracted from the differences between pixels.Stereo image matching estimates the disparity between a rectified image pair, which is of great importance to depth sensing, autonomous driving, and other related tasks.

Problem Definition

- To estimate the disparity images for predicting the depth in an image based on the stereo(left,right) images of cameras which are separated by a fixed distance.Here we use group-wise correlation network which belongs to cross-correlation method.

Functional Requirement

- Feature extraction module should extract the features from left and right images with half dilation setting.
- The concatenation cost volume should concate the left and right features with more parameters than usual.
- The group wise correlation cost volume should group the similar features and provide special parameters for concatenation volume.
- The aggregation network should give four outputs so that it helps in predicting accurate disparity maps.

Functional Requirement

- Input:
  - R1)The stereo image pair should be rectified one.
  - R2)The input image size should be (height x width) .if not crop the images to that size.
- Output:
  - R3)Improve the network such that the disparity images should be real time.
  - R4)Improve the network such that the accurate corresponding points are found in the ill posed regions such as occlusion areas, repeated patterns, textureless regions, and reflective surfaces.
  - R5)Improve the Cost volume network such that less information is lost.

Modules

- Unary Feature Extraction
- Cost Volume Construction
- 3D Aggregation
- Disparity Prediction

- **Unary Feature Extraction**

  For feature extraction, we adopt the ResNet with the half dilation settings and without its spatial pyramid pooling module. We denote the channels of unary features as N_c. All the channels are evenly divided into N_g groups along the channel dimension, and each feature group therefore has N_c/N_g channels. The gth feature group f_l^g, f_r^g consists of the $g * N_c/N_g$, $g * N_c/N_g + 1, \ldots, g * N_c/N_g + (N_c/N_g 1)$th channels of the original feature f_l, f_r.

Modules

- **Cost Volume Construction**

  The cost volume is composed of two parts, a concatenation volume and a group-wise correlation volume.The full correlation provides an efficient way for measuring feature similarities, but it loses much information.The concatenation volume contains no information about the feature similarities. we propose group-wise correlation by combining both.

  The group-wise correlation is then computed as :

$$C_{gwc}(d, x, y, g) = 1/(Nc/Ng)((f_l^g(x, y), f_r^g(x - d, y))$$

Modules

- **3D Aggregation**

  The 3D aggregation network is used to aggregate features from neighboring disparities and pixels to predict refined cost volumes. It consists of a pre-hourglass module and three stacked 3D hourglass networks to regularize the feature volumes. As shown in Figure 1.2, the pre-hourglass module consists of four 3D convolutions with batch normalization and ReLU. Three stacked 3D hourglass networks are followed to refine low-texture ambiguities and occlusion parts by encoder-decoder structures.

Modules

- **Disparity Prediction**

  Two 3D convolutions are employed to output a 1-channel 4D volume, and then the volume is upsampled and converted into a probability volume. For each pixel, we have a Dmax-length vector which contains the probability p for all disparity levels. Then, the disparity estimation d is given by

$$d = \sum_{k=0}^{Dmax-1} k.pk$$

  where k and pk denote a possible disparity level and the corresponding probability The predicted disparity maps from the four output modules are denoted as d_0, d_1, d_2, d_3. The final loss is given by

$$L = \sum_{i=0}^{i=3} \lambda_i.Smooth_{L1}(d_i - d^*)$$

Low level design

- **Model**

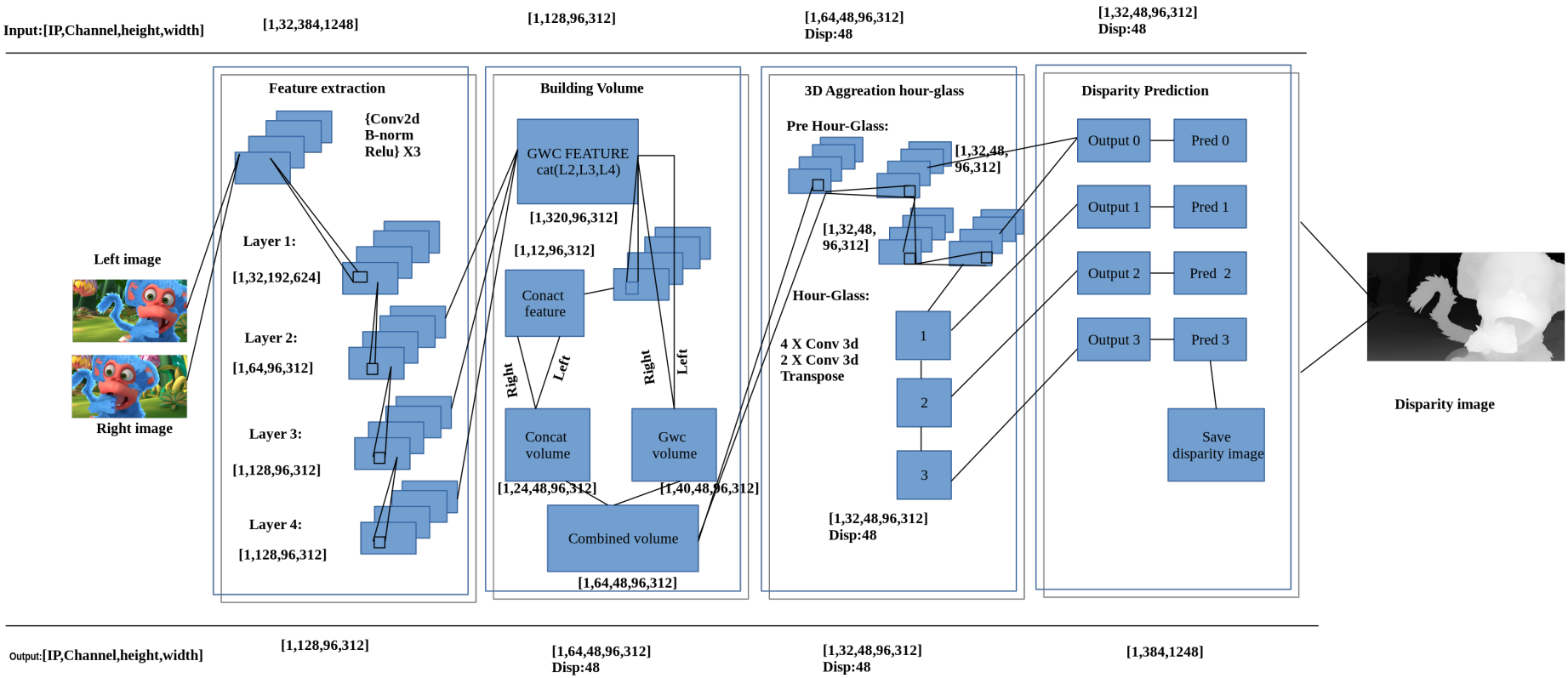

Fig 1.1 : Low Level Diagram

Low level design

- Module1:
  Feature extraction:
  1. The input to the this module are Rectified Stereo Image pair of size 960x540 (Height x Width),with batch size=1 and epoch =300.
  2. The output of this module is a tensor which contains the left and right features of the image.i.e images of size H x W of 320 channels(320 features).
- Module2:
  Cost volume construction::
  1. Input for this module is a 320 channel image pair with size H x W .i.e all the features of left and right image pairs.
  2. Output of this module is a feature map with 64 channels and of size H x W x D.

Low level Design

- Module3::
  3D aggregation Network::
  1. Input this network is the feature map with 64 channels and of size H x W x D.
  2. Output of this module has four feature map with 32 channel and with size H x W x D
- Module4::
  Disparity Prediction::
  1. Input to this module is are four feature maps with size H x W x Z.
  2. Output of this module is the predicted disparity image.

Flow Chart

- **Flowchart**

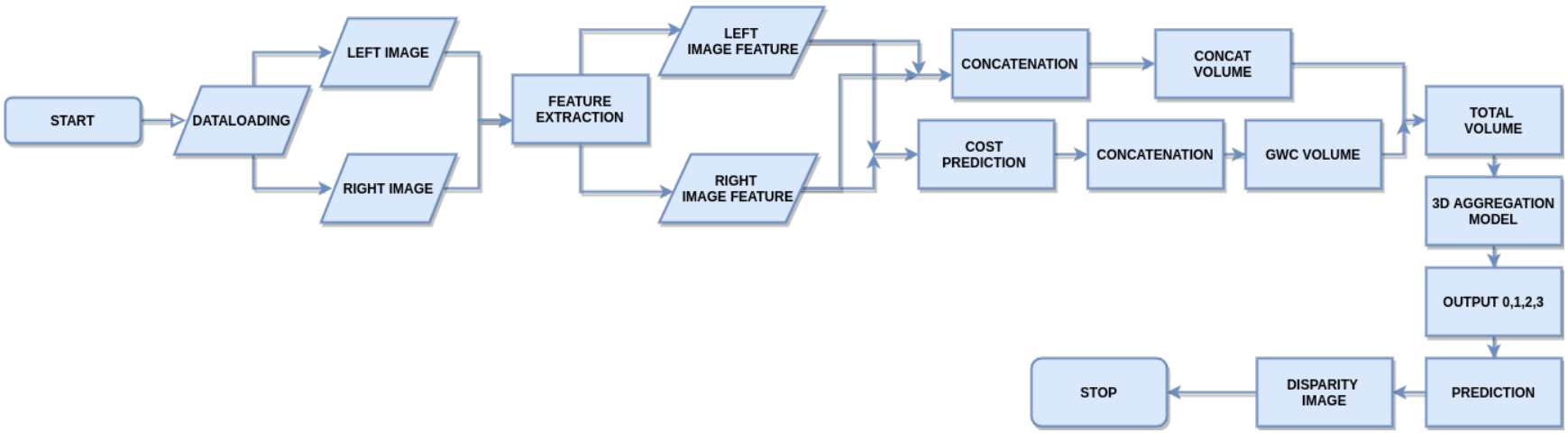

Fig 1.1 : Flow Of The Netwrok

Test Cases

- We evaluated our method on three stereo datasets: Scene Flow, KITTI 2012, and KITTI 2015.
- Scene Flow datasets are a dataset collection of synthetic stereo datasets, consisting of Flyingthings3D, Driving, and Monkaa. The datasets provide 35,454 training and 4,370 testing images of size 960540 with accurate ground-truth disparity maps.
- KITTI 2012 and KITTI 2015 are driving scene datasets. KITTI 2012 provides 194 training and 195 testing images pairs, and KITTI 2015 provides 200 training and 200 testing image pairs. Both datasets provide sparse LIDAR groundtruth disparity for the training images.

Result Analysis

- **Sceneflow dataset**

Fig 2.1 : Left Image

Fig 2.2 : Right Image Result

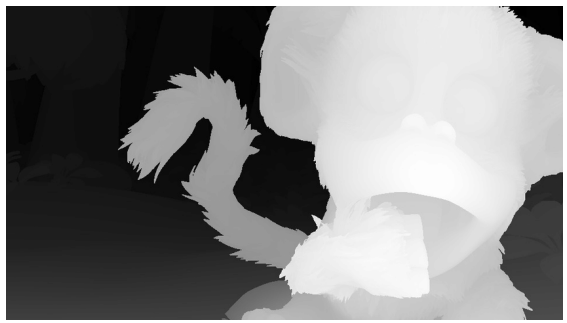

Fig 2.31 : Disparity Result

Result Analysis

- **Kitti Dataset**

Fig 2.32 : Disparity Result

Conclusion

- The groupwise correlation volumes provide good matching features for the 3D aggregation network, which improves the performance and reduces the parameter requirements of the aggregation network.
- To show that when the computational cost is limited, our model achieves larger gain than previous concatenation-volume based stereo networks.
- To imporve the stacked hourglass networks to further improve the performance and reduce the inference time.