INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Literature review on depth estimation using a single image

Vyshna R. K.
vyshnark@gmail.com
Model Engineering College, Kochi, Kerala

Dr. Priya S.
priya@mec.ac.in
Model Engineering College, Kochi, Kerala

ABSTRACT

Depth estimation is an important component of understanding geometric relations within a scene. Different depth estimation techniques using a single image are analyzed in this survey paper. Depth estimation from a single image is often described as an ill-posed and inherently ambiguous problem. Recovering depth information in applications like 3D modeling, robotics, autonomous driving, etc. is more important when no other information such as stereo images, optical flow, or point clouds are unavailable. For the task of depth estimation using single images, learning based methods have shown very promising results.

Keywords— *RGB-D Image, CNN, Multi-scale deep network, Depth estimation, Higher resolution maps*

1. INTRODUCTION

Nowadays depth information is used by almost every vision algorithm. Almost all the state of art algorithms use some kind of combination of the data including depth and extract as much information as they can from it. RGB-D images are pretty common and are used in the Autonomous Driving technology, a significant boost in noticed in the accuracy of the depth estimation techniques.

We can calculate the disparity map using the stereo images, by estimating the distance moved by a particular point in the left and right images. An inverse correlation exists between the shift of the point and the distance of the object from the camera. If larger the point's movement then smaller will be the distance of the object from the camera. Instead of searching for a point in the whole image, a stereo camera assumes the epipolar geometry, its searches for a particular point only along the horizontal-x-axis of the stereo images. To find the correspondence of a point, in the left image, to its position in the right image along the same x-axis, usually with an offset and range. This drastically reduces the space complexity of our algorithm and reduces the multi-dimensional problem to a single dimension. Since there is much prior work on estimating depth based on stereo images or motion, there has been relatively little on depth estimation from a single image. Even though the problem of finding correspondences in the two images sound easy, it is quite difficult to estimate the depth from a single image.

With the available accurate image correspondences, depth can be recovered deterministically in the case of stereo images [14]. But estimating depth from a single image requires the use of monocular depth cues such as line angles and perspective, object sizes, image position, and atmospheric effects. And also, a global view of the scene may be needed to relate these effectively, whereas local disparity is sufficient for stereo images. Depth estimation of a scene from a single image is an easy task for humans but is challenging tasks for computational models to do with higher accuracy. But this problem has attracted a lot of attention in the field of computer vision, since depth maps have a lot of applications, such as augmented reality [10], human-computer interaction [12], segmentation [11] and scene recognition [13]. With the recent employment of Convolutional Neural Networks (CNN) has accelerated the research of the problems. Based on the superior results with the introduction of CNN's, it is apparent learning deep featuresare superior to hand-crafted features.

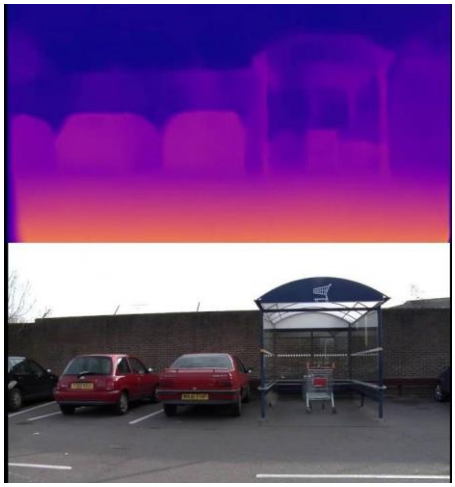

Fig. 1: Example of depth prediction on KITTI [8] dataset

2. PREDICTION FROM A SINGLE IMAGE USING A MULTI-SCALE DEEP NETWORK

This approach directly regresses on the depth using a neural network with various components. One that estimates the global structure of the scene and second that refines it using local information [3]. These component stacks are applied to the original input. But along with it, the coarse networks output is

passed to the fine network as additional first-layer image features. By this, the local network can edit the global prediction to incorporate finer-scale details. The global coarse-scale network is to predict the overall depth map structure. The upper network layers are fully connected, and thus contain the entire image in their field of view. Likewise, the lower and middle layers are designed to combine information from different parts of the image through max-pooling operations. Also, this network takes care of things like vanishing points, object locations, and room alignment. The final output is at 1/4-resolution compared to input. Local fine-scale network edits the coarse prediction it receives to align with local details. It has three convolutional layers with coarse features concatenated to the second block.

In addition to point-wise error, this network is trained using a loss that explicitly accounts for depth relations between pixel locations. Along with the improved qualitative outputs, this system achieves state-of-the-art estimation rates on NYU Depth [5] and KITTI [8] datasets.

3. CLASSIFICATION USING DEEP FULLY CONVOLUTIONAL RESIDUAL NETWORKS

Formulating depth estimation as a pixel-wise classification task. Discretizes the continuous ground-truth depths into several bins and label the bins according to their depth ranges. In the case of CNN, input image is inevitably down sampled due to the repeated combination of max-pooling and striding. In order to handle this, this approach follows the Fully Convolutional Network (FCN) which is successful in dense pixel labeling. Replaces the fully connected layers in conventional CNN architectures with convolutional layers. The model consists of a convolutional layer, a max pooling layer followed by 4 convolution blocks. A building block with linear projection followed by different numbers of building blocks with identity mapping is contained in each convolution block. Here two deep residual network architectures with 101 and 152 layers are considered. It performs a bilinear interpolation on the map to make it the same size with the input image during prediction.

To formulate depth estimation as a classification task, it uses a pixel-wise multinomial logistic loss function. The continuous depth values are uniformly discretized into multiple bins in the log space. Here the label index of a pixel indicates its distance. The labels of typical classification tasks are different from the depth labels. Here each prediction is in the form of probability distribution and this confidence is achieved by formulating depth estimation as classification. In this approach, this is also applied during post-processing via fully connected CRFs. This approach is evaluated on the NYUD2 and KITTI datasets and compared with recent depth estimation methods.

4. UNSUPERVISED MONOCULAR DEPTH ESTIMATION WITH LEFT-RIGHT CONSISTENCY

This approach enables the convolutional neural network to learn to perform single image depth estimation, in the absence of ground truth depth data. It generates disparity images by training the network with an image reconstruction loss. It performs end-to-end unsupervised monocular depth estimation with a novel training loss to enforce left-right consistency inside the network. It introduces new losses to help train the network to predict depth with high accuracy.

This model proposes a function able to reconstruct one image from another. At training time, the images corresponding to the left and right color images from a calibrated stereo pair is given to the network. The network attempts to find dense correspondence when applied to the left image, which helps to reconstruct the right image.

The bilinear sampler is used to generate the predicted image with backward mapping. Using a novel left-right consistency loss Godard et al. [1] train the model to predict both disparity maps. The fully convolutional network consists of an encoder and a decoder along with skip connections. By using a single input in testing, the network still predicts two disparity maps.

Introduces an L1 left-right disparity consistency penalty to ensure coherence between the left and right disparity maps, which attempts to make the left-view disparity map equal to the projected right-view disparity map,

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

The disparity map can now be converted to a depth map, using the camera distance and focal length. This method achieved a 4.935 RMSE score and threshold accuracy of 0.976 on the KITTI dataset.

5. BASED ON FOURIER DOMAIN ANALYSIS

A deep learning algorithm approach for single image depth estimation based on the Fourier frequency domain analysis. The algorithm is based on ResNet-152, which is a very deep network, including 151 convolution layers and 1 fully-connected layer. Out of 50 blocks, last 19 blocks are modified here, which has an additional path for intermediate feature extraction. It uses a two-phase method for training. The first phase includes, train the network after removing the additional feature extraction parts and maintaining the original structure of ResNet-152. In the second phase, it uses the parameters from the first phase but initializes the parameters of the additional feature extraction parts with Gaussian random values. This method helps to facilitate faster training and also to improve the depth estimation performance.

In regression problems, when the Euclidean loss is employed, in general, the network is trained to estimate the depths of distant objects more reliably than those of near ones. In this approach, to overcome this problem, a new loss is introduced, called Depth-Balanced Euclidean (DBE) loss, given by:

$$L_{DBE} = \frac{1}{2N}\sum_{x}\left(g\left(\hat{d}_x\right) - g\left(d_x\right)\right)^2 \tag{2}$$

Where g is a quadratic function for the balancing

$$g\left(d\right) = a_1 d + \frac{a_2}{2}d^2 \tag{3}$$

To reliably estimate shallow depths, as well as deep depths, the network can be trained. It uses NYUV2 depth dataset for training. Also, this approach generates multiple depth map candidates by cropping an input image with various cropping ratios. And to exploit the complementary properties of different depth map candidates, it is combined in the Fourier domain. Thus the proposed algorithm outperforms the conventional algorithms significantly.

6. HIGHER RESOLUTION MAPS WITH ACCURATE OBJECT BOUNDARIES

Most of the existing methods suffer from loss of spatial resolution in the estimated depth maps; including blurry

reconstruction of object boundaries. This approach introduces more accurate estimation with a focus on depth maps with higher spatial resolution. It introduces two improvements to existing approaches. One is about fusing features extracted at different scales, for which it proposes an improved network architecture consisting of four modules. And another approach is about loss functions for measuring inference errors used in training.

It consists of four modules including an Encoder (E), a Decoder (D), a Multi-scale Feature Fusion module (MFF), and a Refinement module (R). The encoder extracts features at multiple scales and the decoder employs four up projection modules, which gradually up-scale the final feature from the encoder while decreasing the number of channels. Four different scale features from the encoder are integrated by MFF module using up-projection and channel-wise concatenation. To merge different information at multiple scales into one the MFF module is used.

Introduces three loss terms, which are used to measure errors in depth, gradients and surface normals, respectively. It also contributes to the improvement of accuracy in a complementary fashion. By including the two improvements, which helped to attain higher accuracy than the current state-of-the-arts, which is given by finer resolution reconstruction.

7. DATASETS

7.1 NYUD-V2

The NYU Depth V2 dataset contains video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. Contains 1449 densely labeled pairs of aligned RGB and depth images, and also 464 scenes taken from 3 cities. Along with it, contains 407,000 unlabeled frames. Inorder to speed up training down-sampling the images are common. And in the case of the raw dataset, it includes raw image and accelerometer dumps from the kinect. The RGB and Depth camera sampling rates usually lies between 20 and 30 FPS.

7.2 KITTI

The KITTI dataset has been created by recording from a moving platform and was introduced in IJRR in 2013. The dataset contains camera images, laser scans, high-precision GPS measurements and IMU accelerations from a combined GPS/IMU system. This dataset place an important role in the development of computer vision and robotic algorithms targeted for autonomous driving. And also most of the experiments are following the Eigen split, instead of utilizing the entire KITTI dataset, which contains 23,488 images from 32 scenes for training and 697 images from 29 scenes for testing.

8. COMPARATIVE STUDY

Estimating depth from single monocular images is a key component in the case of scene understanding. Due to the continuous property of depths, most of the existing algorithms formulate depth estimation as a regression problem.

In the Depth Map Prediction from a Single Image using a Multi-Scale Deep Network by Eigen et al. [3] uses multiscale features for monocular depth estimation. And by joining the information from both global and local views, the model performance is reasonably well. And for this system uses two deep networks, one for estimating the global depth structure, and another network for refining it locally at a finer resolution. And also this model achieves a new state-of-the-art by using NYU Depth and KITTI datasets. Also introduced scale-invariant loss which helps mitigate errors in depth due to scale. This model can improve the performance by incorporating further 3D geometry information, such as surface normals and also by repeated application of successively finer-scaled local networks extending the depth maps to the full original input resolution.

Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks [9] estimates depth from single monocular images by using deep fully convolutional residual network architecture. In this paper, continuous depth values are discretized into different bins and depth estimation is formulated as a discrete classification problem. By using this confidence of a prediction can be easily obtained and can be applied during training which uses information to gain matrices as well as post-processing by using fully-connected CRFs. The performance can be improved by introducing the multi-scale inputs and by upsampling the prediction maps.

In Unsupervised Monocular Depth Estimation with LeftRight Consistency by Godard et al. [1] introduces new inputs. They use left and the right image for training purpose and a single image for testing the network. In testing, from a single image view, it reconstructs the other images disparity map and also constructs the corresponding depth map. They introduce different losses such as matching loss, disparity smoothness loss, and left-right disparity consistency loss. And also this model can generalize to unseen datasets and still able to produce visually plausible depth maps. The performance of this model can be further improved by adding temporal consistency.

In the paper Single-Image, Depth Estimation Based on Fourier Domain Analysis [7] proposes a deep learning algorithm for single image depth estimation based on the CNN and Fourier frequency domain analysis, by introducing a new loss function, called depth balanced Euclidean loss. This loss function trains the network reliably for a wide range of depths. And also it generates multiple depth map candidates by cropping an input image with various cropping ratios and combines them in the Fourier domain.

Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries [6] covers the loss of spatial resolution in the estimated depth maps, for example, a typical symptom is a distorted and blurry reconstruction of object boundaries. The improved network architecture consists of four modules: an encoder, a decoder, a multi-scale feature fusion and a refinement module. And it proposes to use a combined loss of the three loss functions, and it provides significant improvement especially for estimation of fine details and clear boundaries of objects in scenes.

9. CONCLUSION

Learning based methods are showing very promising results in the task of estimating depth from single images. And by analyzing several methods for depth estimation using a single image with various improvements and pitfalls we can conclude that still there are many more effective and efficient methods can be introduced to solve this problem with more precision. And to unfold the issues regarding depth estimation from a single image still, several proposals are available to analyze.

10. REFERENCES

[1] C. Godard, O. Mac Aodha, and G. J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in CVPR, vol. 2, no. 6, 2017, p. 7.

[2] Y. Cao, Z. Wu, and C. Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, pages 23662374, 2014.

[4] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In CVPR, pages 20022011, 2018.

[5] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

[6] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries.arXiv e-prints, page arXiv: 1803.08673, Mar 2018.

[7] Lee, Jae-Han Heo, Minhyeok Kim, Kyung-Rae Kim, Chang-Su. (2018). Single-Image Depth Estimation Based on Fourier Domain Analysis. 330-339. 10.1109/CVPR.2018.00042.

[8] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, Sparsity invariant cnns, in International Conference on 3D Vision (3DV), 2017.

[9] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks.CoRR, abs/1605.02305, 2016.

[10] W. Lee, N. Park, and W. Woo. Depth-assisted real-time 3d object detection for augmented reality. ICAT, 2:126132, 2011.

[11] Y. Lu, J. Zhou, J. Wang, J. Chen, K. Smith, C. Wilder, and S. Wang. Curve-structure segmentation from depth maps: A CNN-based approach and its application to exploring cultural heritage objects. AAAI, 2018.

[12] Z. Ren, J. Meng, and J. Yuan. Depth camera based hand gesture recognition and its applications in human-computer-interaction. 2011 8th International Conference on Information, Communications and Signal Processing, pages 15, 2011.

[13] S. Zia, B. Yksel, D. Yret, and Y. Yemez. Rgb-d object recognition using deep convolutional neural networks. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 887894. IEEE, 2017.

[14] R. I. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

