Unsupervised stereo image depth estimation of midan recovery structure

lixianwei, leidian
School of computer science, Shanghai University, Shanghai 200444, China
leidian@staff.shu.edu.cn

Abstract. The stereo image depth estimation is an important field in computer vision. At present, most of the researches on depth estimation based on deep learning make use of the mathematical characteristics of the image, such as left-right consistency and pixel distribution characteristics. In this paper, we pay attention to the relationship between the grid representation, which is a distance representation mode of biological system, and the depth representation. Inspired by the concept of Fourier decomposition, we propose a median recovery network structure. The Experiment shows that this network structure has a certain effect on depth estimation.

1. Introduction
Image depth estimation has become a very important and popular research direction in computer vision. In the early stage, image depth estimation mostly uses artificial features, such as image hidden points and shadow. Because most of these features have special requirements for the image, it is difficult to make breakthroughs in image depth estimation. With the development of computer vision, more and more artificial features have been proposed, such as scale invariant feature transformation (SIFT), hierarchical gradient direction histogram (Phog) and other different artificial features[1]. Most of the artificial features mentioned above focus on the local information of the image instead of the global information. Therefore, sexena and other researchers have introduced the probability method and constructed different probability graph models such as conditional random field (CRF) and Markov random field (MRF) [2]. The model considers both global information and long-distance information, so the problem is transformed into a learning problem under random fields. Using machine learning tool, many important achievements have been made.

With the development of deep learning, especially convolutional neural network (CNN), more and more researchers in the field of image depth estimation also apply CNN to image depth estimation, and obtain many very important research results. At present, more and more innovation of deep estimation is based on deep learning.

Animals, especially humans, perform very well in depth estimation. Relevant researchers have also used many natural characteristics of human body, such as attention mechanism and the consistency of left and right depth of binocular image. In this paper, based on the grid cell representation theory discovered by Daniel Bush et al. [3] and Andrea banio et al. [4] on how distance is represented in animals through grid cells, we combine the grid cell theory with Fourier decomposition to improve the depth estimation network.
2. Related work

There are many methods used for image depth estimation. According to the dataset used for training, they can be roughly divided into three categories: monocular image (monocular image dataset) depth estimation, binocular image (stereo image pair dataset) depth estimation, and video image (multi-view image dataset) depth estimation. Binocular image depth estimation is favored by many researchers, and has become the key research directions in the field of depth estimation due to the complete theory and sufficient dataset.

2.1 Monocular image depth estimation

Egen et al. [5] was the first one to apply the CNN to depth estimation. They used fine-coarse two-stage method to estimate the depth of monocular images in a supervised way, which achieve great result than the previous research. In the following research, Egen et al. [6] combined depth estimation with surface normal prediction and semantic segmentation, this three tasks is similar, and finally achieved good results. Laina et al. [7] greatly increased the depth of the network. As a result, the resolution of network output depth map reach half of input image. The main contribution of Laina is that they used full convolution network which eliminates the limitation of the previous network for the size of the input image. At the same time, the speed of upconvolution module is improved in their research.

Aleksei grigorev et al. [8] introduced the LSTM network into the previous CNN depth estimation network, which used for processing the feature map obtained by CNN, and combined the context features to make the depth map more smooth and accurate. Cao et al. [9] discretized the depth map and converted the depth estimation into a classification task by using CRF and probability model, which effectively combined the probability model with the depth estimation task. It also brought new problems, such as how to discrete the depth map is proper.

Later, Huan Fu et al. [10] used SID(spacing increasing discretization) to discrete the depth map. According to the distance distribution in the depth map, the depth map is discreted and the task is transformed into an ordered regression task. In this paper, they improved the full image encoder and proposed the ASPP module which bring super visual field. Jin Han Lee et al. [11] proposed the local planar guidance module based on the local smoothing hypothesis The module predicts the depth maps with different resolutions at different scales, and finally obtains the final high-resolution depth maps through the joint prediction of multi-scale depth maps. In this paper, ASPP module (as shown in Fig. 1a) is used to solve the problems of low resolution feature map caused by pooling layer.

2.2 Stereo image depth estimation

Different from monocular image depth estimation, binocular image depth estimation is mostly based on the stereo geometry relationship between two images captured by binocular camera. After the improvement maded by deep learning, Juyuan Xie et al. [12] began to use it in 2D to 3D image conversion. They proposed a network called deep3D, which added the diaparity map obtained by multi-scale deconvolution together, and used a convolution and softmax activation to obtain different scale probability disparity maps. The synthesis right image through the disparity map and the left image.

After that, Ravi Garg et al. [13] used the depth convolution network to obtain the disparity map through the left image, and finally synthesis the left image from the right image and the parallax map. Through the optical flow error, they can compare the synthesized image with the original image. Through the accurate disparity map and the mathematical relationship between the disparity map and the depth map, the final depth map can be obtained. Later, Alex Kendall et al. [14] constructed a cost volume network by using depth feature map and geometric structure knowledge of binocular image. They used 3D convolution to integrate context information on cost volume and the final disparity value was obtained through a fully differentiable soft argmin operation. Godard et al. [15] further improved the binocular image depth estimation method by reconstructing the left and right images at the same time, and proposed the left and right disparity consistent loss, which makes the left and right disparity map consistent to improve the performance of depth prediction. On the basis of Godard's work, Wang Benzhang et al. [16] added a deep improvement network to the end, and proposed a perceptual loss
function which encourage the input image and the reconstructed image to be best matched, so as to generate high-quality disparity map.

2.3 Grid cell representation
Daniel Bush et al. [17] explored how caries animals use patterns of different scales to represent space. Caries animals can use specific periodic representation to represent a very large space through a set of patterns of different scales (as shown in Fig. 1b). Inspired by the concept of Fourier decomposition, they also explore how to reproduce specific displacement from a set of periodic representations, and the possibility of navigation based on grid representation of olfactory cells in carious animals was explored. In the following study, Andrea banio [18] used the grid representation mode for the navigation of artificial agents. The author integrated the path through the RNN network, and finally got the grid cell like representation, which prove the grid pattern can accurately represent the distance.

![Fig1. (a) ASPP module[11] (b)Grid cell module[17]](image)

3. Method
Compared with monocular image depth estimation, the stereo image image depth estimation which based on the geometric relationship between binocular images has the following advantages:

1) There are accurate geometric relationship between stereo image. Through the left (right) image $I_1(x)$ and diaparity map $D(x)$, the right (left) reconstructed image $I_2(x + D(x))$ can be recovered. By forcing the reconstructed image consistent with the original image, the disparity map can be obtained. The depth map can obtain through the accurate mathematical formula $d = bf / (dis)$, where $b$ is the distance between binocular cameras and $f$ is the focal length of the camera.

2) Compared with monocular image supervised method, binocular image data set is easier to collect because it does not need depth truth image.

3) The final accuracy of monocular supervised method is limited because of the missing points of the groundtruth depth map is difficult to be accurate. The geometric relationship between binocular images is objective. Theoretically, the diaparity map obtained by stereo method can be very accurate.

3.1 Midan recovery theory
Depth estimation is a global scalar prediction task, and the depth is a certain scalar value between the target point and camera. In the grid cell model, the distance scalar value is decomposed into several periodic quantities represented by different scale components; In Fourier decomposition, a scalar value can also be decomposed into the accumulated values of multiple frequency components. In the above two theories, the closer the restored variables are to the values before decomposition when more component values on different scales (frequencies) are obtained. One image can be decomposed to many brightness variation with different frequency, amplitude and phase by Fourier transform of image. We can restore one specific image with different components. The more components we get, the more accurate we can recover. Based on the above theory, we putforward one consistent structure in depth estimation, based on following theory:
1) **Average depth map:** if the depth map of an image is predicted at multiple different scales, such as four different scales. For the adjacent disparity map, the depth map corresponding to the smaller scale predicted depth map is the approximate average depth map corresponding to the adjacent large scale depth map, that is, the main frequency component obtained by Fourier decomposition of an accurate depth map. As shown in Fig. 2, the depth map before and after upconv is the average depth map of the current scale depth map.

2) **Small differences:** there are only small differences in the fineness between the depth maps on different scales on adjacent scales. After Taylor expansion of the depth maps, the depth maps on different scales only lack the small details on different times (these differences are infinitely small in theory). The formation of these specific differences is due to the small field of vision of the front network layer in the encoder network, these layers can only get local information. The local information mainly contains detail, such as texture and object boundary. For the latter network layer, its view field is larger due to the existence of a large number of pooling layers and strides in conv block, and the feature map contains global structure information. Therefore, on a smaller scale, most information received by the network is the global structure information, and the local details has been filtered by the front network layer of the encoder network, so only the coarse scale depth map can be predicted on the smaller scale. For the two adjacent scales, the feature map received on the larger scale network only has more details, so there are only small detail differences between the adjacent scale depth maps.

3) **Cumulative restoration method:** Based on the idea of average depth map and small difference, the following recovery methods are proposed for depth map prediction: Firstly, the depth map should be predicted on multiple scales; Secondly, at the minimum scale, the global depth map on the coarsest scale is obtained by using convolution network on the feature map through reconstructing the corresponding left(right) image as the corresponding right(left) image. Thirdly, the depth map on the lower scale should be upsampled as the base depth map of the current scale depth map. Forthly, on the basis of the base depth map, the corresponding layer feature map of the encoder network is combined with the feature map of the current decoder layer by using skip-connection method. Fifthly, the Aspp module and conv layer is used to obtain the additional depth detail of the current scale size, which is called small difference map. Finally, we can restore the depth map of current scale through the base depth map and small difference map. The corresponding midan recovery structure in every scale is shown in Fig2.
The median recovery structure proposed in this paper combines all the depth maps of different scales together. So, the improvement of depth maps of any scale will improve the following depth maps.

3.2 network architecture

The network in this paper is constructed according to the restoration ideas proposed above. The network is an encoder-decoder network, which is composed of two parts: the encoder feature extractor and the decoder prediction part. The encoder part consists of the encoder network commonly used in image classification and recognition. Inspired by Godard et al. [13], this paper adopts an unsupervised way to reconstruct the right image from the left image, and reconstruct the left image from the right image at the same time.

Traditional encoder networks are mainly used for image recognition and classification. These tasks pay much attention to the outline and global information of the image, but pay less attention to the details. Therefore, these encoder networks use a large number of pooling layers to achieve two purposes: 1. Expand the network field of each scale. 2. Filter some unnecessary local details. Depth estimation requires both field and local detail information. Therefore, most of the current network use skip connection to obtain the information from the previous layer of encoder network in order to supplement the lack local details of the decoder network.

Huan Fu et al. [10] proposed ASPP module, which can expand the field of vision without reducing the resolution of feature map. It can indirectly combine different length information by using different rate values. Through ASPP, we can get different length information without losing resolution in order to obtain better depth map. The global information and the current scale details can be combined to obtain better depth map at different scale. The network is shown in Fig3.

The network contains the principle structure of median recovery proposed in this paper. It uses ASPP to obtain different range of information, and predicts depth map on the minimum scale. The depth map on the minimum scale is used as the base depth map. By combining the base depth map with small difference map, we can predict corresponding disparity map at different scales. There are contact between different scale disparity maps. The improvement of disparity map at any scale will improve the following disparity map.

![Fig3. network architecture](image-url)
3.3 Loss function
In this paper, the network predicts the depth at four different scales. Therefore, we need to add loss functions at four different scales. Because the structure of four scales is consistent, the loss function is the same on four different scales.

In this paper, the loss function is composed of two parts: image matching loss and disparity map smoothing loss. The image matching loss is mainly used to compare the structural similarity and brightness similarity between the input image and the reconstructed image, so we choose the SSIM error:

\[
C_{im} = \frac{1}{N} \sum_{i,j} 1 - SSIM(I_{i,j}, \tilde{I}_{i,j})
\]  

(1)

Most of the current depth maps have good performance in global structures. However, in the object boundary, the depth map obtained by hardware equipment or network has the same shortcomings, such as depth discontinuity and depth value missing. In order to make up for the lack of depth and other shortcomings, we add the smooth loss to the total loss. By adjusting the weight value before the smooth loss and compare loss, we can compromise between effect and time:

\[
C_{ds} = \frac{1}{N} \sum_{i,j} \left| \partial_x d_{i,j}^l \right| e^{\frac{1}{2} \left| x_{i,j} - y_{i,j} \right|} + \left| \partial_y d_{i,j}^l \right| e^{\frac{1}{2} \left| x_{i,j} - y_{i,j} \right|}
\]  

(2)

By combining the above two loss functions, we can get the total loss as this:

\[
C = \alpha_1 (e_{ap}^l + e_{ap}^r) + \alpha_2 (e_{ds}^l + e_{ds}^r)
\]  

(3)

4. Experiment and result

4.1 The KITTI dataset
In this paper, we use stereo image to predict depth map. At present, many depth datasets only provide monocular images and corresponding depth maps, which is not suitable for our network. After comparison, Kitti dataset is selected as the training dataset in this paper. Kitti dataset is a large urban outdoor dataset. Due to the limitation of time, equipment and personnel, we only use a small part of the calibrated binocular data as the total training dataset. The training dataset contains 300 images, including 25 different outdoor scenes, and the verification dataset contains 100 images, including 9 different scenes. In addition, 40 additional images are set aside as final test samples to verify the performance of the network.

4.2 Implementation details
The network is deployed based on keras, using tensorflow as backend, including 43 million training parameters. The equipment we use is Dell workstation, which uses a 1080ti GPU for training, and the total training times are 50 epochs. The network accepts left and right binocular images with the size of 300x1200 as network training input. According to Odena et al. [19], we modified the network, using bilinear interpolation layer without learning parameters to replace traditional conv layer. The network contains two user-defined layers. One layer is used for downsampling the input image and the other bilinear interpolation layer is used for reconstructing corresponding image. These two layers do not contain any trainable parameters. Therefore, there is no difference between the network and the network composed of standard keras layer. We can build the network with standard keras.

In the training stage, the proportion of matching loss function and smoothness loss function are set to 1 and 0.5 respectively. The upsampling disparity map, small difference map and the sampled image are combined to predict the disparity map. Finally, the bilinear interpolation module is used to reconstruct the image structure.

| Method   | Supervised | Dataset  | Abs Rel | Sq Rel | RMSE  | RMSE Log |
|----------|------------|----------|---------|--------|-------|----------|
| Eigen[3] | Yes        | KITTI    | 0.214   | 1.605  | 6.563 | 0.292    |
| Eigen[4] | Yes        | KITTI    | 0.203   | 1.548  | 6.307 | 0.292    |
4.3 Result and discussion

Figure 4 shows the prediction results of the network in the Kitti dataset. The input images are randomly selected from the images in the test dataset. The image shown in Fig. 4 (b) is the corresponding network prediction depth map. It can be seen from the results that the network performs well in terms of the overall structure, the boundaries of different objects, and the overall far and near characteristics. However, the hierarchical performance of far objects is poor. For example, the depth of the overall prediction of the far car is the same level, and the depth difference is not predicted.

![Fig4. (a) the input left image (b) the depth image (after highlight)](image)

There are two reasons for this problem: the network and minimum representation distance. Firstly, we use left(right) image to construct right(left) images to obtain disparity map in order to obtain depth map in this paper. In the distant objects, the disparity between stereo image is small, which makes it difficult to learn the difference between stereo image; secondly, the maximum image depth map is only 255. The depth range of Kitti dataset is 0-80m, so the minimum accuracy is 0.31m which is not enough to represent sufficient hierarchy.

From the results of Fig. 4 (b), it can be seen that the hierarchy of distance is not shown on the surface with sparse texture (flat road in Figure 4). The reconstruction and matching of sparse texture surface is a difficult problem in the academic field, and we will explore it in the future.

5. Conclusion

In this paper, we combine binocular depth prediction with grid representation, and finally propose a multi-scale median recovery idea and corresponding median recovery structure. Based on this idea, we construct a corresponding multi-scale stereo disparity map prediction network. By using the mathematical relationship between the disparity map and the depth map, we can obtain the depth map indirectly. Based on the Kitti dataset, the corresponding multi-scale network is trained, and the verification experiment is carried out. The results show that the proposed median restoration idea and the corresponding stereo disparity prediction network have good results in stereo disparity map prediction.

In the following work, the network will be further improved, including expanding the training samples of the network, improving the loss function of the network, and further verifying the network capacity upper limit and rationality of the current structure. The ASPP module used in the network in this paper is very consistent with the idea of median recovery proposed in this paper. The relationship between the two will be discussed, in order to improve the idea proposed in this paper.

References

[1] Karsch K, Liu C, Kang S B. Depth Extraction from Video Using Non-parametric
[1] Sampling[C]. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012, pp. 775-788.

[2] Saxena A, Min S, Ng A Y. Learning 3-D Scene Structure from a Single Still Image[C]. IEEE International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8.

[3] Bush D, Barry C, Manson D, et al. Using Grid Cells for Navigation[J]. Neuron, 2015, 87(3):507-520.

[4] Banino, A., Barry, C., Uria, B. et al. Vector-based navigation using grid-like representations in artificial agents[J]. Nature 557, 429–433 (2018).

[5] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single Image using a Multi-scale Deep Network[J]. In NIPS. 2014.

[6] Eigen D, Fergus R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture[C]. IEEE International Conference on Computer Vision. IEEE, 2015.

[7] Laina I, Rupprecht C, Belagiannis V, et al. Deeper Depth Prediction with Fully Convolutional Residual Networks[J]. 2016.

[8] Aleksei Grigorev. Depth estimation from single monocular images using deep hybrid network[J]. Multimed Tools Appl (2017) 76:18585–18604.

[9] Yuanzhouhan Cao. Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks[J]. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. 2018(28):3174-3182.

[10] Huan Fu, Mingming Gong, Chaohui Wang. Deep Ordinal Regression Network for Monocular Depth Estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[11] Lee J H, Han M K, Ko D W, et al. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation[J]. 2019.

[12] Xie, J., Girshick, R., & Farhadi, A. Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks[C]. European Conference on Computer Vision. Springer, Cham. 2016.

[13] Garg R, Vijay Kumar B G, Reid I. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue[J]. 2016.

[14] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-End Learning of Geometry and Context for Deep Stereo Regression[C]. IEEE International Conference on Computer Vision. IEEE, 2017.

[15] Godard, Clément, Mac Aodha O, Brostow G J. Unsupervised Monocular Depth Estimation with Left-Right Consistency[C]. Computer Vision & Pattern Recognition. IEEE, 2017.

[16] W. Benzhang, F. Yiliu, F. Huini and H. Liu. Unsupervised Stereo Depth Estimation Refined by Perceptual Loss[C]. 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), 2018, pp. 1-6.

[17] Karsch K, Liu C, Kang S B. Depth Extraction from Video Using Non-parametric Sampling[C]. European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012, pp. 775-788.

[18] Saxena A, Min S, Ng A Y. Learning 3-D Scene Structure from a Single Still Image[C]. IEEE International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8.

[19] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checker board artifacts[J]. Distill, 2016.