Monocular Image Depth Calculation Based on Convolution Neural Network

Shi Zhibo

Zhengzhou University of Industry Technology, Department of Information Engineering, Zhengzhou

Keywords: Depth calculation; Depth estimation; Convolution neural network; Multi-scale; DenseNet

Abstract: The acquisition of depth information is a very important part of scene analysis, which is mainly divided into two methods: sensor acquisition and image processing. The technology of sensor technology has very high demands on the circumference, so the image processing is a more general method. The traditional method uses binocular stereo calibration to get the depth by geometric relation, but it is still restrained because of the circumstance. Therefore, as the closest to the actual situation of the square method, monocular image depth estimator has a very large research value; For this reason, a method for depth estimation of monocular image based on DenseNet is proposed, which is based on multi-scale convolutional neural network. The DenseNet structure is added to optimize the feature acquisition process by using the characteristics of DenseNet strong feature transfer and feature reuse. The results show that the average relative error is 0.119. The root mean square error is 0.547. And the logarithmic mean space error is 0.052.

1. Introduction

With the rapid development of artificial intelligence, people of all kinds of intelligent products (such as driverless cars, medical robots, patrol robots) came into being. In the course of its work, it is necessary to make a decision according to the outer boundary environment element. The general realization method is to use computer vision technology to perceive the environment in three-dimensional structure, realize three-dimensional reconstruction, and make decision.

Therefore, 3D scene analysis is one of the hottest and most important research topics in artificial intelligence. The important basis of 3D scene analysis is the acquisition of depth information, and the one-eye image is one of the important methods to obtain depth information. At present, there are two methods to obtain the depth of the scene: hardware reality and software implementation. The hardware-in-the-loop method is a technique of teleportation, such as the 3D body sense camera Kinect developed by Microsoft, which uses the ToF principle to code the invisible light and to vary the time between the strong and weak distance light lines, according to the reflection of light, the special point is real time. The workload of algorithm development is low. However, the resolution of the image is too low, which is only suitable for small-range indoor environment measurement.

The method of software realization is to obtain depth information through image processing. At present, the most common method is the binocular standing body calibration method[1]. The parallax is obtained by matching the same feature points in the left and right images. On this basis, the structure light method is put forward, which is based on coded light source, and the main action is to match the optical source on the image. Because of the low requirement of setting quantity and environment condition for monocular image, it is the most flexible method to obtain depth information of monocular image and apply scene. Using monocular image to obtain the depth of information is a disease-like question, one image can theoretically be applied to the unrestricted depth map. Therefore, the traditional method to obtain depth information of monocular image is based on the movement of object and the change of focal length of camera[2]. SFM (Strict-from-Motion) is one of the methods[3] used by SFM to obtain the depth information of the object in the scene.

With the development of deep learning in recent years, Convolutional Neural Networks (CNN) play a more and more important role in image processing and speech recognition. CNN does not need human intervention in feature extraction, result classification and so on, which greatly

DOI: 10.25236/icmcs.2019.053

improves the universality of the model. Therefore, the researchers began to use CNN widely to study the depth information estimation of monocular images. In view of the excellent performance of CRF in semantic segmentation, Liu et al.[4] put forward a method of applying CRF to depth estimation of monocular image, which is different from the one-dimensional and two-dimensional potential function method of CRF. By optimizing the potential function of the training coefficient, we can train the results of the training, but the super-pixel mentioned in this paper still needs to be divided manually. By contrast, Eigen et al.[5] proposed a method of depth learning based on multi-scale network junction structure, which can obtain the final output of the whole-office and the whole-office samples respectively by two scales. Training directly on the original image and obtain pixel-level depth information results. On the basis of Eigen et al., Laina et al.[6], a network model using ResNet[7] is proposed, which uses the ResNet feature to forward transfer the high-efficiency characteristics, and combines the deeper and more complex network junction structure to effectively improve the fruit fineness. Go-dard et al.[8] proposed a method of unsupervised learning based on the left-right graph. The principle of this method is similar to binocular setting, and the left-right graph is used to obtain the depth map, and the left-right graph is used to pre-measure the fruit of the left graph. Self-coding and self-decoding are realized, but the above-mentioned method can improve the result precision. In order to solve the above problems, a kind of monocular image depth estimation method based on depth study is proposed. According to the input RGB image, the depth information of each pixel in the graph is obtained directly. The main new point of this paper is: Firstly, a new CNN structure is proposed, which can extract and fuse the image on three scales, and optimize the output results. Secondly, DenseNet[9] is integrated into the network structure, which enhances the forward propagation of the features, alleviates the problem of eliminating the loss of the deep network ladder degree, and strengthens the special weight, realizes the multi-level comprehensive and efficient utilization, and also reduces the parameter number at the same time. Finally, the experiments show that the multi-scale network structure and the addition of DenseNet can effectively improve the accuracy of the output depth map.

2. Network Model

2.1 Model Overview

In order to further study the depth estimation method of monocular image, a multi-scale CNN network model based on DenseNet is proposed in this paper. Firstly, the network structure is divided into three scales, in which the first scale (Scale1) is the largest input image and the third scale (Scale3) is the smallest. The first scale takes the global rough sampling of the image features, and the output fruit is the same as the input image of the second scale (Scale2).The input image of Scale2 is combined with the input image of Scale1 on the basis of the original data set. Scale2 pays more attention to the local information collection in the image and optimizes the global and rough results on the previous scale to obtain the results with more local features. Similarly, the input of Scale3 is a combination of the original data set and the output of Scale2, which improves the resolution of the depth map and achieves high resolution output. The specific network structure is shown in Figure 1.

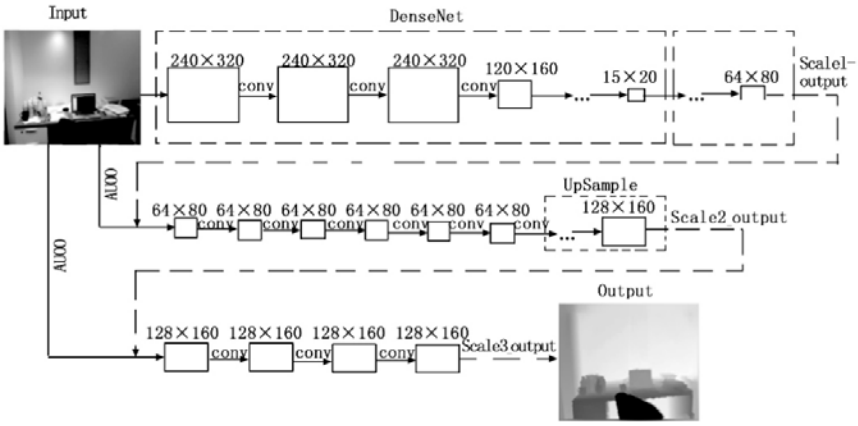

Figure 1 Network Model Structure

2.2 Multi-scale network structure

The main object of Scale1 is the sampling of the image in the whole bureau, extracting image features at the whole bureau level. The network consists of the DenseNet module, the upper sampling module and the convolution layer. Scale1 's input image is 240 × 320. First, the image is collected through two convolution layers. Then, the convolution layer (called 3 × 3 convolution layer) with convolution kernel 3 × 3 and step 2 is used to sample the image downwards instead of the traditional pool layer. The result is input as DenseNet module. After the image is passed through the DenseNet module, the output result is sampled up by the upper sampling module, and the final image output size is 64 × 80, which is the same as the input size of Scale2.Since Scale1 has the largest input image size, Scale1 has a wider field of vision, and the special features collected are the most abundant and the most original.

Scale2 expands the number of input samples and the size of the input image is 64 × 80 by combining the shrinking data set with the result output of Scale1 through the concatenated layer. Compared with Scale1, Scale2 takes into account the correlation between pixel depth and peripheral pixel depth information in the image. Scale2 network concatenation structure is composed of convolution layer and upper sampling module, and the convolution layer is responsible for rich-featured information sampling. The upper sampling module is responsible for the delivery of proper discrimination rate. In order to replace the traditional 5 × 5,7 × 7 convolution nuclei with 3 × 3 convolution nuclei, more non-linear factors exist between 3 × 3 convolution nuclei and large convolution nuclei under the condition of the same receptive field. At the same time, the realization of 3 × 3 convolution kernel can significantly reduce the number of parameters and improve the training speed. Finally, Scale2 outputs a depth image of 128 × 160 in size, the same size as Scale3 input.

Scale3 combines the scaled dataset with the result output of Scale2 through the concatenatelayer, and the network collation is similar to the Scale2 class.Scale3 is responsible for further optimizing the image to send out and raise the high discrimination rate, can be connected to the space, detailed output results.

2.3 DenseNet module

With the development of in-depth learning, various CNN network models have been proposed, from the beginning of AlexNet, VGGNet, to the future of ResNet and so on. In 2017, DenseNet was proposed by Liu Zhuang and others. The network structure has the following characteristics: First, DenseNet effectively alleviates the problem of gradient dispersion caused by the deep network, and DenseNet can obtain the loss function of each layer in front of it, effectively strengthen the feature forward propagation, so it can train the deeper network. Secondly, compared with ResNet's square transmission characteristic, DenseNet uses a splicing method to transfer the layers in front of it into each other in order to increase the rate of transmission efficiency significantly, and its non-linear transformation is shown as formula 1. Thirdly, DenseNet effectively reduces network parameters and improves network performance from the point of feature reuse. For example, the ResNet-based network model prevents over-fitting through random discard layers, the table is not necessary for all layers, and there is a large amount of redundancy in the network network, which makes the wave cost into the volume of operation. In general, DenseNet requires only half the number of ResNet parameters for the same pre-test accuracy.

$$x_l = H_l([x_0, x_1 ....., x_{l-1}]) \tag{1}$$

DenseNet is composed of DenseNet DenseNet, Transformed Layer, Pool Layer, Convolution Layer and Full Connection Layer. DenseBlock is a dense connected high-way module, which is a set that is transmitted by BottleneckLayer at each layer. Bottle-neckLayer consists of convolution layers, activation functions, normalization functions, and Drop-out. TransentionLayer is responsible for connecting adjacent DenseBlock to further normalize DenseBlock's output, Dropout, and so on.

In this paper, the DenseNet module is composed of 4 DenseBlock and 3 Transformed Layer, and it has been improved in the original structure of DenseNet, which is the main knot of Scale1 network. Firstly, for the image initialization part, the 3 × 3 convolution layer is used to replace the 7 × 7 convolution layer and the pool layer, and the large convolution kernel is replaced by the small convolution kernel. Secondly, according to the number of output channels of DenseBlock, the number of output channels of each layer is determined to be 6,12,48,32 from top to bottom.Then,3 × 3 convolution layer is used to replace the pool layer for TranssitionLayer. Finally, the global pooling layer and the fully connected layer at the end of the DenseNet network structure are removed and replaced with a better 1 × 1 convolution layer.

2.4 Upper sampling module

The main purpose of upper sampling is to amplify the original image and obtain higher resolution output. The method of interpolation is basically used to enlarge the image. Based on the original image pixels, the interpolation algorithm is used to insert new pixels between pixels. In the depth learning model, the first step is to expand the image resolution to 2 times, fill the pixels without data with 0, and then use 5 × 5 convolution layer to process the extended image to realize interpolation. Because there are a lot of 0 values, there are a lot of useless computation. In this paper, the upper sampling module divides the 5 × 5 convolution kernel into 4 convolution nuclei of size 3 × 3,2 × 3,3 × 2,2 × 2[6], which operate directly on the original drawing, skipping the processing of 0 value. Then the 4 output fruit are intersected and misconnected, and the fruit is lost.At the same time, on the basis of this, the Resu number is added, and the integral function number is added to the convolution layer, and the sampling module is combined to improve the efficiency of upper sampling.

3. Results and Analysis

3.1 Experimental setting

In order to verify the optimization effect of this model, this paper chooses to train and test the model on NYU DepthV2 official data set. NYU DepthV2 is a video frame sequence of indoor scenes captured by Microsoft's Kinect camera, which consists of images corresponding to 1499's depth information and RGB pixels. The data set contains 464 scenes in three cities, which are divided into 26 scene types and more than 1000 objects. The training set and test set are divided into 249:215. The original 480 × 640 RGB graph and the depth graph are sampled down to 240 × 320 as the model input. The default pixels of the depth information in the image are ignored by preprocessing. According to the official classification, through 49 scenes for verification set,200 scenes as training set, training completed in the official 654 standard quasi-verification map on the model test. In this paper, we extend the data set by processing the image of the training map, such as the progressive reduction, the rotation of the target plane, the horizontal inversion, and the change of color and contrast.

Avoid over-fitting model, improve generalization ability. The experimental equipment is a display card for the TeslaK40C service operator, making use of the TensorFlow framework. Network training using stochastic training gradient descent optimization module

Type parameter, the specific hyperparameters are as follows: batch size (batchsize) is 8, maximum iterative wheel number (maxpoch) is 1000. Learning rate (learningrate) is 0.The learning rate declines by 90% for every 10 iterations until the network converges. The training time of the model is about 72 hours, and the forward process of CNN is about 0.0 per graph.06s. The whole module type, the pre-measurement time of each graph is about 0.23 s. The loss curve during training is shown in Figure 2. In this paper, the experimental results of the model are compared with those of training on the NYU DepthV2 dataset, and the evaluation results of the usual weights are used[10]:

(1) Average relative error (REL)

$$\mathrm{Re}\, l = \frac{1}{T}\sum_{i}\frac{|\, y_i^* - y_i \,|}{y_i^*} \tag{2}$$

(2) Root means square error (RMS)

$$RMS = \sqrt{\frac{1}{T}\sum_{i}(y_i^* - y)^2} \tag{3}$$

(3) Logarithmic spatial mean error (aeragelog10error, log10):

$$\log 10 = \frac{1}{T}\sum_{i}|\log y_i^* - \log y_i \;| \tag{4}$$

(4) Accuracy:

The percentage of pixels satisfied as a total pixel. $\max(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*}) = \delta < threadhold$ In the above formula, T is the sum of the number of pixels in the test image, with the predicted value and the true depth value of the model, which are pixel i, respectively.

The learning process of the whole training model is:

Step 1: Enter the RGB image $L = (x_0, x_1 ....., x_{n-1})$ and the true depth chart, and expand the data set.

Step 2: RGB image after the network model to get output y, compared with the real depth value y_i^*, to obtain $d_i = | y - y^* |$.

Step 3: Calculate the loss function based on d_i size, and update the parameters.

Step 4: Repetition step 2,3, if the final model runs when it reaches the upper limit of the retracting bar or the supernumerary number.

3.2 Experimental results

It is shown in Table 1 that the resulting fruit will be compared with make3D, Eigen, and Laina's training methods to output the fruit into the row. First of all, we can see that the output result of CNN convolution network model is obviously better than that of traditional geometric hypothesis. This is mainly due to the strong ability of CNN structure in image processing. CNN structure can extract enough feature information from the image. There is no need for manual intervention in special extraction and classification. In the network model using CNN structure, the output results are slightly better than those of Eigen and Laina models, mainly due to DenseNet's network characteristics and multi-scale network connection.

Table 1 Comparison of Experimental Results on the NYU DepthV2 Dataset

Method	Error (the smaller the better)			Accuracy (the bigger the better)		
	Rel	RMS (RMS)	Log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Make3D	0.349	1.214		0.449	0.745	0.897
Eigenetc (Eigen etc)	0.158	0.641		0.769	0.950	0.988
Laina etc	0.127	0.573	0.055	0.811	0.953	0.988
This text	0.119	0.547	0.052	0.821	0.958	0.988

4. Conclusion

In view of the problem of depth pre-measurement of monocular image, a multi-scale network complex junction pattern based on deep degree learning is proposed, which is based on DenseNet's strong characteristic forward propagation, feature reuse to reduce the number of parameters, global pre-measurement provided by multi-scale network collaterals and the special combination of local pre-test fruit knots. It can improve the depth of single-eye image, and the result is better than that of other CNN network models. But at present this model is still under the supervision of the governor under the article training, need to bring up the true depth of information, because this pair of data sets mentioned a high demand. The author is trying to build an unsupervised training model, to increase the scope of application of the model. At the same time, the model is directly trained by RGB graph to get the depth result. The next step is to consider integrating the traditional method of depth acquisition into the model, for example, to improve further the prediction accuracy with dual-target principles.

References

[1] ZHANG Zhengyou, MA Songde. Computer vision[M],Beijing: science press,1998.

[2] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[A]//Advances in neural information processing systems[C]. 2006: 1161-1168.

[3] Szeliski R.Structure from motion[J]. Computer Science, 2011:303-314

[4] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5162-5170.

[5] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in neural information processing systems. 2014: 2366-2374.

[6] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 239-248.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 270-279.

[9] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[10] LI Yaoyu, WANG Hongmin, Structured Deep Learning Based Depth Estimation from a Monocular Image[J], robot, 2017,39(6):812-81