Surface Height Map Estimation from a Single Image Using Convolutional Neural Networks

Xiaowei Zhou†, Guoqiang Zhong†, Lin Qi†, Junyu Dong*†,Tuan D. Pham§ and Jianzhou Mao‡
†Ocean University of China, §Linkoping University, ‡Macau University of Science and Technology
Email: dongjunyu@ouc.edu.cn.

ABSTRACT

Surface height map estimation is an important task in high-resolution 3D reconstruction. This task differs from general scene depth estimation in the fact that surface height maps contain more high frequency information or fine details. Existing methods based on radar or other equipments can be used for large-scale scene depth recovery, but might fail in small-scale surface height map estimation. Although some methods are available for surface height reconstruction based on multiple images, e.g. photometric stereo, height map estimation directly from a single image is still a challenging issue. In this paper, we present a novel method based on convolutional neural networks (CNNs) for estimating the height map from a single image, without any equipments or extra prior knowledge of the image contents. Experimental results based on procedural and real texture datasets show the proposed algorithm is effective and reliable.

Keywords: Surface height map estimation, surface reconstruction, CNNs.

1. INTRODUCTION

Surface height map estimation is very important for high-resolution 3D reconstruction. A surface height map is a raster image that stores height values, such as geometric height of a bumpy surface [1]. We can display the height map as luma of a grayscale image (Fig.1 (a)). The height map can be used to render the surface (Fig.1 (b)) by using LuxRender. Traditionally, large-scale scene depth or geometry information can be captured with the aid of active equipments, such as radar, laser scan and Kinect [2]. Liu et al. proposed a method based on the results from an Interferometric SAR (InSAR) analysis to estimate the height map of high-rise buildings [3]. However, these methods might fail in estimating the height map of a textured surface, which contains high frequency information or fine details. On the other hand, some methods are available for estimating surface height map from multiple images, e.g. photometric stereo [4], or from a single image with prior knowledge [5]. In order to estimate surface height map, photometric stereo was introduced in [4]. Han et al. estimated detailed shape of objects from single RGB-D image through estimating light model and shape-from-shading approach [5]. Nevertheless, estimating surface height map from a single image without other equipments or extra prior knowledge of the image contents is still a challenging and open issue.

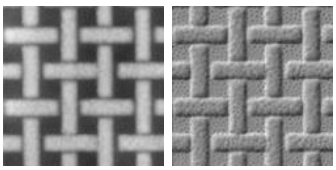

Figure 1. (a) Surface height map; (b) rendered image.

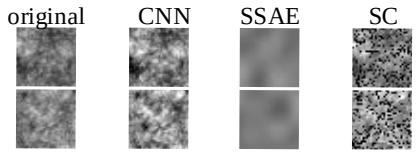

Figure 2. The original surface height maps and surface height maps estimated by CNN (our), SSAE and SC.

In recent years, deep learning has achieved state-of-the-art results in many fields. Liu et al. proposed a method based on deep CNNs [6] and CRF to estimate depth from single image [7]. David et al. predicted depth, surface normals and semantic labels by using a common multi-scale convolutional architecture [8]. These methods focus on predicting general large-scale scene depth, such as meeting room and dining hall. Our task differs from general scene depth estimation in the fact that surface height maps contain more high frequency information or fine details. In contrast to previous work, we proposed a method based on convolutional neural networks to predict a surface height map from a single image without extra equipments or prior knowledge. We have performed extensive experiments on a procedural texture dataset [9] and a real texture dataset PerTex [10], to verify the effectiveness of the proposed method. Experimental results show that our method is effective to estimate the surface height map from a rendered image. The

contribution of the proposed method is a deep-learning based model for estimating surface height map with fine details from a single image without any other auxiliary equipments or prior knowledge.

2. MODELS

In this section, we introduce our proposed models for surface height map estimation in detail. Our work refers to estimating surface height map from a single image. Given an image, we can obtain the corresponding surface height map through our trained model. It can be regarded as a pixel-to-pixel prediction problem. We designed a regression model based on convolutional neural networks during prediction process. The proposed model includes convolutional layer, full connection layer and sigmoid nonlinear transformation layer. In order to design such a regression model, the loss layer has been set to the Euclidean loss layer. The definition of the Euclidean loss is: $Loss = \frac{1}{2N}\sum_{n=1}^{N}\|y_n - \hat{y}_n\|_2^2$, where N is the number of image samples; y_n is the predicted surface height map; and $\hat{y}_n$ is the ground truth. So the $Loss$ is the average loss of all the images. After each convolutional layer and full connection layer, we add a sigmoid nonlinear transformation layer, since the scope of our input and output is between 0 and 1.

Specifically, we have different neural networks with different number of layers to handle different size of images. For small size (less than 70×70 we recommend) images, we designed a small scale convolutional neural network inspired by LeNet [11]. Our network is not similar to the conventional convolutional neural networks. It only contains the convolutional layer and full connection layer, without the pooling layer. Our experimental results have demonstrated that better results can be produced without the pooling layer.

For large size (greater than 70×70 we recommend) images, we designed a large scale convolutional neural network inspired by GoogleNet [12]. The proposed network architecture is modified from GoogleNet. The network architecture can be found in the open source project (https://github.com/zhouxiaowei1120/HeightEstimation). GoogleNet is a 22 layers deep network and there are three softmax loss layers. We extracted all the network layers before the first softmax loss layer as the foundation of our model. Then we made some modification to this part of the model. We removed the dropout layers to reserve more information and removed one full connection layer to reduce memory consumption. Then replaced the softmax loss layer with Euclidean loss layer. At the same time, we replaced all the Relu nonlinear transformation layers with sigmoid nonlinear transformation layer, since the data scope was between 0 and 1.

Different scales of networks can accept different size of input images. The small scale network consisting of few network layers is more efficient. Furthermore, if large size images were input into small scale network, inferior experimental results would be produced. That is because a small scale convolutional neural network is too simple to have the learning ability to represent large images. If small size images were input into large scale network, it would be inefficient for our task and hard to train the model.

3. EXPERIMENTAL RESULTS

In this section, we show the experimental results of our proposed model on a procedural texture dataset [9], a real texture dataset PerTex [10] and natural images. These two texture datasets include both surface height maps and rendered images, which are essential for training and testing the proposed methods. We obtained promising results by using our model. Unless stated, the pixel values in our experiments were normalized to $[0, 1]$.

3.1 Results on Procedural Texture Dataset

3.1.1 Experiments on Small Images

We carried out extensive experiments to estimate surface height map from input images. In addition to convolutional neural networks, we also compared sparse coding [13] and stacked sparse Auto-encoder [14] for small images. We found that the method based on convolutional neural networks produced the better results.

We first tested the sparse coding (SC) method for surface height map estimation. Assuming that the input vector $\mathbf{y}$ could be represented as a linear combination of a set of basis vectors ϕ_i: $\mathbf{y} = \sum_{i=1}^{k} \alpha_i \phi_i$, where α_i is the code; ϕ_i is also called dictionary. We used 20,000 original surface height maps or rendered images in the dataset as the dictionary for surface height maps (ϕ_1) and rendered images (ϕ_2). We employed sparse coding to obtain the code $\mathbf{x}$ for the rendered images. Meanwhile, the codes for rendered images and those of the corresponding surface height map are treated as the same ones. In this way, we can obtain the surface height map by $\phi_1 \times \mathbf{x}$ for each rendered image. For stacked sparse

Auto-encoder (SSAE), we used a four-layer Auto-encoder. In the training process, rendered images were used as input; corresponding surface height maps were employed in the reconstruction layer in each Auto-encoder. For testing, the input of the model was rendered image, and the output of the model was surface height map. The number of hidden layer units of the four Auto-encoder was 2048-800-2000-1024. The sparsity parameter is set to 0.01 and the weight decay is set to 10^{-4}. For the small scale convolutional neural networks, we implement our models on the platform of Caffe [15]. Training and testing are performed on a standard desktop with an NVIDIA Titan X GPU. The base learning rate is set to 0.01; learning rate policy is set to "inv". The specific network parameters can be found in the open source project.

We compared the results of the above three models on the procedural texture model fractal (one-over-fBeta-noise) [9]. The size of images in the fractal (one-over-fBeta-noise) texture model is 512×512 and the number of images is 100. We cropped each image into 32×32 small images. So the total number of cropped surface height maps in fractal model is 25,600. After cropping the images, we rendered the 25,600 cropped surface height maps to produce the corresponding rendered images by using the same render method as [9]. We took 25,500 images for training and 100 images for testing. The estimating result is shown in Fig.2. We can easily find that the CNN produces the best result.

3.1.2 Experiments on Large Images

For large images, we did experiments on a procedural texture dataset we called it Fractal dataset by using large scale network. The network parameters are in the open source project. All the images in this dataset were generated by ourself using LuxRender, according to the fractal (one-over-fBeta-noise) texture generation model [9]. We performed experiments on this dataset, which includes 23,749 images (23,700 images for training and 49 images for testing) whose size is 128×128. The estimated surface height maps are shown in Fig.3. We can easily see that the estimated surface height maps and the original ones are almost same. The proposed model is effective for estimating surface height maps.

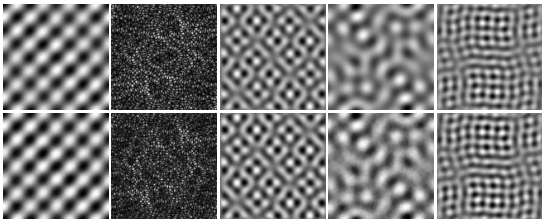

Figure 3. The first row are the original surface height maps and the second row are the estimated surface height maps.

Table Ⅰ The MSE of estimated surface height maps and estimated rendered images.

	height map		rendered image	
image size	32×32	128×128	32×32	128×128
Our model	**0.0016**	0.017608	**0.0208**	0.0554
SSAE	0.0074	-	0.1034	-
SC	0.0318	-	0.0368	-

We computed the mean square error (MSE) of the test samples to measure the errors between the estimated surface height map and the ground truth. MSE is defined as: $\frac{1}{2n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2$. Where n is the number of the pixels; p_i is the estimated value; $\hat{p}_i$ is the ground truth. Table Ⅰ shows the MSE of small size and large size images by using our model. In the table, the first and the second column are the MSE between the estimated surface height map and the ground truth. The third column and the fourth column are the MSE between the estimated rendered images and the original rendered images. From Table Ⅰ, we can see that the error for one pixel is very small compared with the data scope. Experimental results show our model is better than sparse coding and sparse stacked Auto-encoder. ``-" represents that the experiments can not be carried out on 128×128 images by SSAE and SC.

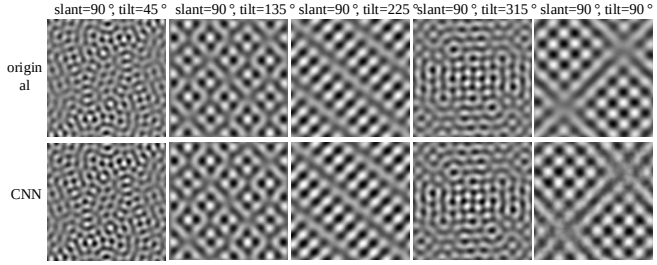

Figure 4. The original surface height maps and the estimated surface height maps from five rendered images under five different illumination conditions.

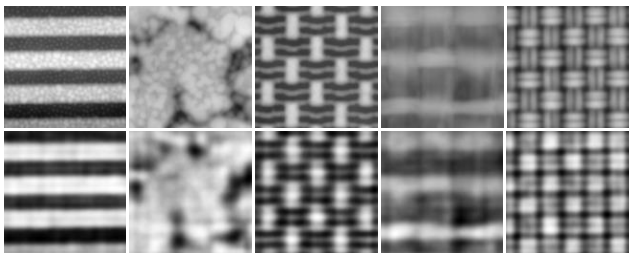

Figure 5. The first row are the original surface height maps and the second row are the estimated surface height maps.

The above experiments are carried out on the procedural texture dataset whose images are rendered under a fixed illumination condition. In order to illustrate the generalization of our model, we carried out experiments on a new procedural texture dataset. This new procedural texture dataset whose images were rendered under different illumination conditions was constructed on the foundation of Fractal dataset. We chose 15,100 surface height maps from the Fractal dataset. The chosen surface height maps were rendered by LuxRender as [9] under four different illumination conditions ($slant = 90^{\circ}, tilt = 45^{\circ}$; $slant = 90^{\circ}, tilt = 135^{\circ}$; $slant = 90^{\circ}, tilt = 225^{\circ}$; $slant = 90^{\circ}, tilt = 315^{\circ}$) to get 64,000 rendered images. The new procedural texture dataset consisted of 60,000 images in the training set and 400 images in the testing set. As shown in Fig.4, the experimental results generated by proposed model (CNN) on the new procedural texture dataset are remarkable. Noted that the last column result is obtained by directly inputting the rendered image under a new illumination condition ($slant = 90^{\circ}, tilt = 90^{\circ}$, not in the training set) to the trained model.

3.2 Results on Real Texture Dataset

We also did experiments on a real-texture dataset PerTex [10] to validate the effectiveness of our model. The 334 texture images in this dataset are all real textures, whose surface height map size is 1024×1024. For sake of same settings as above, we cropped the rendered images and the corresponding surface height maps of the PerTex dataset into size of 128×128. The cropped images were divided into training set and testing set. We randomly chose 21,120 images to construct the training set and the rest 256 images were used to construct the testing set. Then these cropped rendered texture images in the training set and the corresponding surface height maps were used to train the large scale CNN (Network settings are same as above). Once the training process finished, we could input the rendered images in the testing set into our trained model to estimate their surface height maps. The estimated surface height maps and the original surface height maps are shown in Fig.5. As we can see from Fig.5, the two type images are very similar. And the MSE of PerTex test dataset between the estimated surface height map and the ground truth is 0.003342. Compared with data scope [0,1], the MSE is very small. Thus, our model is effective to estimate the surface height maps of real textures.

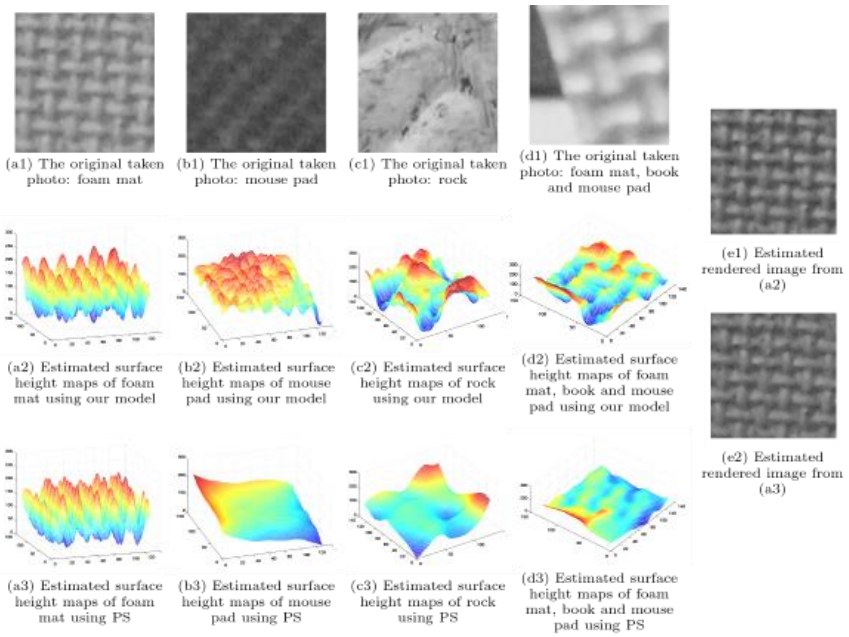

Figure 6. Figures (a2), (a3) are the reconstruction results from (a1) by using different models and so do figures (b), (c) and (d). Figures (e1), (e2) are rendered under the same condition, they are very similar to the original image (a1).

3.3 Results on Natural Images

To validate the availability of our model in practical application, we did experiments on natural images. The natural images were taken from a foam mat, a mouse pad, a rock and a book under different spot light sources by using Canon EOS 550D camera. We compared the experimental results of our model with the classical 3D surface reconstruction method photometric stereo (PS) [4]. These results are shown in Fig.6. Since there is no ground truth of natural images, we compare the results of two models visually. We can easily see that the reconstruction result of our model is

comparable with PS; however, it should be noted that our result is estimated from a single image by using trained model on texture dataset while the result of photometric stereo is produced form several images. Moreover, rich details are recovered by our results and the results of photometric stereo are smoother on account of its enforcing smooth constraint.

4. CONCLUSION AND FUTURE WORK

We presented a method based on convolutional neural networks for estimating surface height map from a single image. The proposed method can estimate the surface height map without any other equipments or extra information. Experimental results show that our method is more effective for estimating surface height map than sparse coding and sparse stacked Auto-encoder. We obtained desirable experimental results on a procedural texture images dataset and a real texture dataset PerTex. We believe that given sufficient training samples, surface height maps of natural images captured under different circumstances can also be well recovered using our model. Maybe we can use some methods [16] to pre-process natural images then to estimate the surface height maps of the natural images.

Acknowledgments: This work was supported by International Science & Technology Cooperation Program of China (ISTCP) (No.2014DFA10410), China Postdoctoral Science Foundation (No.2014M551963), National Natural Science Foundation of China (NSFC) (No.61501417, 41576011, 61271405, 61403353), Open Project Program of the National Laboratory of Pattern Recognition (NLPR) and the Fundamental Research Funds for the Central Universities of China.

REFERENCES

[1] C., Li, and Z. Chen. "Design and implementation of OGRE-based game scene editor software." CiSE, 1-4, (2010).

[2] C. Noguchi, J. Katto, and K. Ohyama, “Improvement of height estimation of low birth weight infants, newborns and infants image processing system using kinect.” GCCE, 221–222, (2014).

[3] W. Liu., K. Suzuki, and F. Yamazaki, “Height estimation for high-rise buildings based on insar analysis.” JURSE, 1-4, (2015).

[4] R. J. Woodham, “Photometric method for determining surface orientation from multiple images,” Optical engineering, vol. 19, no. 1, 191 139–191 139, (1980).

[5] Y. Han, J. Y. Lee, and I. S. Kweon, “High quality shape from a single rgb-d image under uncalibrated natural illumination,” ICCV, 1617–1624, (2013).

[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324, (1998).

[7] F. Liu, C. Shen, and G. Lin. "Deep convolutional neural fields for depth estimation from a single image." CVPR. 5162-5170, (2015).

[8] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” ICCV, 650–2658, (2015).

[9] J. Liu, J. Dong, X. Cai., L. Qi., and M. Chantler, “Visual perception of procedural textures: Identifying perceptual dimensions and predicting generation models,” PloS one, vol. 10, no. 6, p. e0130335, (2015).

[10] F. Halley, “Perceptually relevant browsing environments for large texture databases,” Ph.D. dissertation, Heriot-Watt University, (2012).

[11] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard et al., “Learning algorithms for classification: A comparison on handwritten digit recognition,” Neural networks: the statistical mechanics perspective, vol. 261, 276, (1995).

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” arXiv preprint arXiv:1409.4842, (2014).

[13] M. Yang, L. Zhang, J. Yang, and D. Zhang, “Robust sparse coding for face recognition”. CVPR, 2011. pp. 625–632.

[14] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, “Stacked denoising Auto-encoders: Learning useful representations in a deep network with a local denoising criterion,” The Journal of Machine Learning Research, vol. 11, 3371–3408, (2010).

[15] Y. J., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” arXiv preprint arXiv:1408.5093, (2014).

[16] Deng, L., “Automatic segmentation of solar granulations based on Morphology technique”, Proceeding of the 11th World Congress on Intelligent Control and Automation, 3364-3368 (2014).