EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2018)
R. Sablatnig and M. Wimmer (Editors)

Short Paper

3D Reconstruction and Transparent Visualization of Indonesian Cultural Heritage from a Single Image

Jiao Pan[1], Liang Li[2], Hiroshi Yamaguchi[3], Kyoko Hasegawa[2], Fadjar I. Thufail[4], Bramantara[5] and Satoshi Tanaka[2]

[1]Graduate School of Information Science and Engineering, Ritsumeikan University
[2]College of Information Science and Engineering, Ritsumeikan University
[3]Nara National Research Institute for Cultural Properties
[4]Research Center for Regional Resources, Indonesian Institute of Sciences (P2SDR-LIPI), Jakarta, Indonesia
[5]Borobudur Conservation Office, Magelang, Jawa Tengah, Indonesia

Abstract

Herein, we propose a method for three-dimensional (3D) reconstruction of cultural heritage based on deep learning, which we apply to the reliefs of the Buddhist temple heritage of Borobudur Temple, in Indonesia. Some parts of the Borobudur reliefs have been hidden by stone walls and are not visible following the reinforcements during the Dutch rule. Today, only gray-scale photos of those hidden parts are displayed in the Borobudur Museum. First, we reconstruct 3D point clouds of the hidden reliefs from these photos and predict the pixel-wise depth information for each of them using a deep neural network model. We then apply our stochastic point-based rendering mechanism to produce a high-quality visualization of the reconstructed point clouds. We have achieved promising visualization results that provide us with an intuitive understanding of the valuable relief heritage that is no longer visible to ordinary visitors.

CCS Concepts

•***Applied computing*** → *Digital libraries and archives;* •***Computing methodologies*** → *Neural networks;* •***Human-centered computing*** → *Visualization;*

1. Introduction

Borobudur is a UNESCO World Heritage Site and the largest Buddhist temple in the world. Borobudur comprises approximately 2,672 individual bas-reliefs (1,460 narrative and 1,212 decorative panels), distributed at the hidden foot and the five square platforms. These reliefs can be divided into five sections based on the different independent stories they tell. The temple, which has high cultural value, has been restored and its foot encasement was reinstalled owing to safety concerns. During the restoration, the reliefs of the hidden foot were covered by stones, only grayscale photos photographed in 1890 have remained and are displayed in the Borobudur Museum. Today, only the southeast corner of the hidden foot is revealed and visible to visitors. In this work, we reconstruct the hidden reliefs into point clouds from their monocular grayscale photos. Our reconstruction method is based on a depth prediction neural network which maps intensity or color measurement to depth values. The monocular images and corresponding depth maps of the visible parts of the Borobudur reliefs are used to train the deep neural network. Then, three-dimensional (3D) point clouds are reconstructed from the remaining photos, and the depth map is predicted by the deep neural network model. Furthermore, owing to the Borobudur temple's complex internal structure, we apply our stochastic point-based rendering mechanism to implement transparent visualizations of the reconstructed point clouds.

2. Related Work

With rapid advancements in computer-vision algorithms, it is possible to efficiently and flexibly reconstruct cultural heritage using laser scan data [PMJh14] or defective cultural heritage objects [GSP*14, HS17, LZT*11]. However, many cultural heritage objects, similar to Borobudur reliefs, are no longer available to acquire 3D information from owing to irreversible damage. For image-based reconstruction, the majority of available methods use multiple images to reconstruct 3D models [KL12, KDD*14, IHD*13]. However, only a single monocular photo per object remains prevalent in many cases. Hence, a reconstruction method from a single image is urgently required. Depth estimation from a single image is an ill-posed problem as it is inherently ambiguous to map intensity or color measurement to a depth value. Hand-crafted features and probabilistic graphical models are mainly used to tackle the monocular depth estimation problems in classic methods [SCN05]. Recently, many studies using deep learning have achieved remarkable advances in depth estimation. Eigen et al. [EPF14] were the first to use a convolutional neural network (CNN) to perform depth estimation and produce a promising result over an indoor dataset. As

DOI: 10.2312/gch.20181363

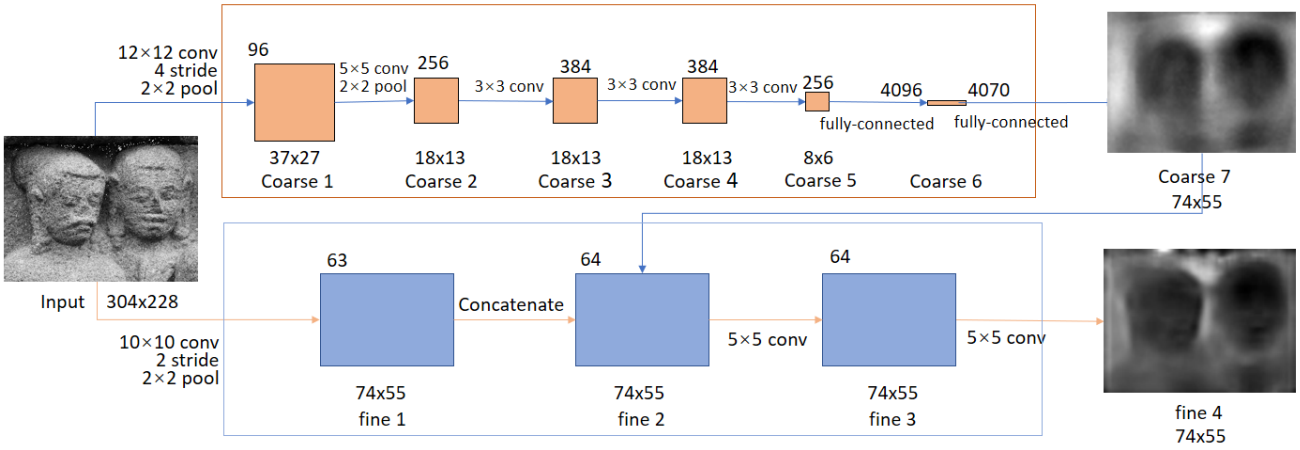

Figure 1: *Network Structure of the proposed method: the orange part at the top represents the global coarse-scale network, and the blue part at the bottom represents the local fine-scale network.*

regards the transparent visualization method for large-scale point clouds, the pioneering approach suffers from a large computational cost due to the depth-sorting process involved. In our previous work, we proposed stochastic point-based rendering, which enables precise and interactive-speed transparent rendering [THS*12]. This method achieves accurate depth feel by employing a stochastic algorithm without depth sorting. We have successfully applied this method to several types of large-scale laser-scanned point clouds of 3D cultural objects [TUH*14, THO*16] and proved its wide applicability.

3. Method

3.1. Network Structure

As shown in Figure 1, the neural network employed for depth prediction comprises two sub-networks: a global coarse-scale network that handles low image resolution and a local fine-scale network that handles high image resolution. The original input image is simultaneously imported by these two networks. The global coarse-scale network predicts a coarse depth map at a global level, i.e., a coarse output with cues such as the object location and an approximate figure shape. The result of the global coarse-scale network is passed to the local fine-scale network and fused with the intermediate results during the convolution. Thus, a refined output is predicted by the latter network. Thus, the coarse output is refined with local details and the final depth map comprises both global and local information.

The global coarse-scale network comprises five convolutional layers with two max-pooling layers to extract feature maps. The feature maps are converted to two fully-connected layers containing the entire image in their field of view. Then, the last fully-connected layer is reshaped into a coarse depth map as the output of the coarse-scale network, in which the value of each pixel is equal to the value of each weight in the last layer. The local fine-scale network comprises four convolutional layers and one max-pooling layer. Following the first convolutional layer and a pooling stage, the output of the global coarse-scale network is imported as a feature map and fused with the original feature maps. Note that the size of the feature maps output by the first part of fine-scale network is identical to those output by the last part of fine-scale network by design. This size is maintained through zero-padded convolutions of all the layers in the local fine-scale network. Figure 1 shows the kernel size and the number of feature maps of each layer. Note that a center crop is maintained following the max-pooling layers in the proposed method. All hidden layers use the rectified linear activations except the two output layers in both networks that use linear activation. To avoid overfitting, we apply a dropout layer after both the coarse 6 and fine 3 parts in Figure 1.

For depth estimation, the question is how to measure the relationships between points in the scene without considering the absolute global scales. Here, we apply a scale-invariant error as a loss function following Eigen' s work [EPF14]. For a predicted depth map y and the ground truth y^*, each with n pixels indexed by i, the per-sample training loss is set as follows:

$$L(y,y^*) = \frac{1}{n}\sum_i d_i^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2 \quad (1)$$

$$d_i = \log y_i - \log y_i^* \quad (2)$$

Note that in Equation 1, the element-wise error is reduced by setting λ=0, and the scale-invariant error is reduced by setting λ=1. When λ=0.5, the function produces absolute-scale predictions, thereby slightly improving the qualitative output.

3.2. Stochastic Point-based Rendering

Here, we briefly review our transparent visualization method, the stochastic point-based rendering [THO*16]. First, multiple subgroups of point clouds are prepared, each of which describes the surface equivalently but is statistically independent. Each subgroup should have the same point-density distribution. Here, we denote

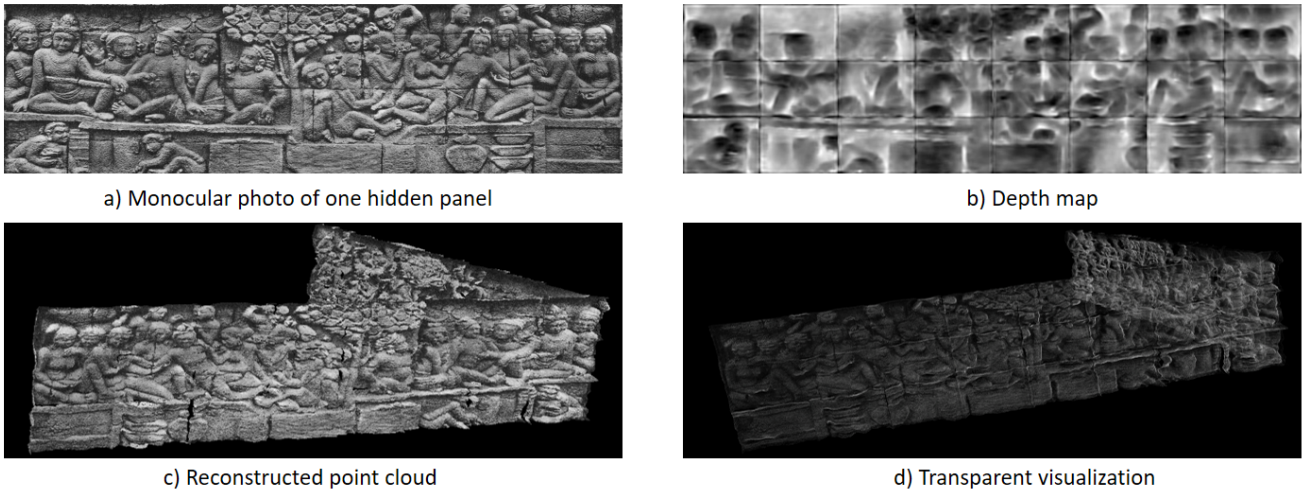

Figure 2: *Main results for one hidden relief of Borobudur: (a) monocular photo of one hidden panel, (b) predicted depth map, (c) reconstructed point cloud and (d) transparent visualization result.*

the number of subgroups by L. Then, for each subgroup, its constituent 3D points are projected onto the image plane to create an intermediate image. In the projection process, we consider the point occlusion per pixel. A total of L intermediate images are obtained. Finally, the L intermediate images are averaged to create the final transparent image. For the number of points, n, the area of the local surface segment S, the point sectional area s, and surface opacity α in each local surface segment adopts the following value:

$$\alpha = 1 - \left(1 - \frac{s}{S}\right)^n \tag{3}$$

In this method, L is available as the image-quality parameter because it reflects the number of the averaged intermediate images. By tuning the local number of points, n, we can control the local surface opacity α according to Equation 3.

4. Experiments

The original dataset used to train the deep network comprises 5,687 pairs of image patches and corresponding depth maps from four monocular photos of four panels of the visible Borobudur reliefs. We simply divided the dataset into training data and test data by a ratio of 75/25. In our case, the four panels contain only 85 human figures and several decorative objects. As the training data is limited in our case, efforts were made to avoid overfitting of the deep neural network. We augmented the training data with several transformations: rotation, flips, adding noise and blurring. After the augmentation, our training dataset contained 31,339 images and their corresponding labels. However, more original data is needed in our case, this is discussed in the future work.

The network was implemented in TensorFlow and all experiments were run on an NVIDIA GTX 1080Ti GPU with 12GB memory. The probability of dropout layers was set to 0.5. The model was trained using an Adam optimizer with the following learning rates: 0.001 for the global coarse-scale network and 0.01 for the local fine-scale network. Our model was fine-tuned on the NYU Depth dataset which comprises 464 indoor scenes, shot as video sequences using a Microsoft Kinect camera. The original size of the frames was 640×480 pixels and we down-sampled it to half the resolution and reshaped it to 304×228 pixels to fit our model. In the pre-training stage, we trained our model with a batch size of 50 for approximately 40 epochs. After the pre-training stage, we trained our model on our relief dataset with a batch size of 50 for approximately 20 epochs. We use the trained model to predict the depth map from the monocular photos of the hidden reliefs. Then, the corresponding point clouds were reconstructed based on the monocular photos and the depth maps. Note that following this, the number of 3D points was the same as the number of pixels of the monocular photo. Furthermore, we apply our stochastic point-based rendering method to the reconstructed point clouds of hidden reliefs to create 3D see-through images.

5. Results

We compared our result with Eigen' s [EPF14] on our relief dataset as shown in Table 1 which proves that our method provides promising results in general for the depth prediction task. The error metrics we used are identical to those in previous works [LSP14]. As our dataset is extremely small, the depth prediction result achieved herein can be improved. While the local fine-scale network seems to exhibit no improvement in the results according to the error

Table 1: *Comparison between Eigen' s and the present results.*

	Coarse-Eigen' s	Refine-Eigen' s	Coarse-Our' s	Refined-Our' s
threshold δ <1.25	0.647	0.612	**0.689**	0.624
threshold δ <1.25	**0.922**	0.906	0.917	0.911
threshold δ <1.25	**0.985**	0.981	0.985	0.983
abs relative	1.046	1.326	**0.872**	1.085
sqr relative	0.716	0.750	**0.715**	0.725
RMSE	10.02	10.16	**10.01**	10.05
RMSE(log)	**0.105**	0.107	0.114	0.113

Figure 3: *Details of the 3D reconstruction results (from left to right): monocular photo patch, depth prediction map, and the reconstructed result.*

metrics, its effect is obviously shown in the depth maps through the sharper and clearer boundaries thus obtained(Figure 1). Figure 2 shows the original monocular photo of one hidden panel of the Borobudur reliefs, the depth prediction map, the reconstruction result and the corresponding transparent visualization result. The gaps in the reconstructed point cloud are caused by the patch-wise training of our network. Improving the accuracy of the depth prediction or implementing post-processing techniques can eliminate the gaps, we will explore this elimination in future work. To clearly demonstrate the details of our reconstruction results, the example results of qualitative patches are shown in Figure 3.

As the temple is approximately a square building, the full-view of these reliefs cannot be viewed at a certain point in case of opaque visualization owing to the existence of corners. Hence, we reconstructed two old photos of the hidden reliefs which represent two panels of the valuable reliefs. Although the reliefs in these photos do not actually cover the two sides of a corner, to explain the contribution of our transparent visualization method, i.e., the stochastic point-based rendering, we simulated the reliefs under the assumption that they do cover the two sides.

6. Conclusions

In this study, we reconstructed 3D point clouds of the hidden parts of reliefs from the Borobudur temple from a single monocular photo. Our reconstruction method achieves a promising visualization result providing a proper 3D understanding of the valuable relief-type heritage that is no longer visible to ordinary visitors. We also applied our stochastic point-based rendering mechanism to the reconstructed point clouds and achieved their see-through imaging with accurate depth feel. We believe that our work provides an efficient way to protect relief-type cultural heritage. We have also demonstrated that our mechanism can be successfully applied to the Borobudur temple reliefs.

In future work, we plan to collect more training data from the visible parts of the Borobudur reliefs using a 3D camera to gain more accurate depth map prediction. An automatic point-cloud-stitching method can also be explored to fill the gaps in the manually-linked point cloud data. Furthermore, we will consider other models to improve the depth prediction results achieved here.

References

[EPF14] EIGEN D., PUHRSCH C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems* (Cambridge, 2014), Neural Information Processing Systems, pp. 2366–2374. 1, 2, 3

[GSP*14] GREGOR R., SIPIRAN I., PAPAIOANNOU G., SCHRECK T., ANDREADIS A., MAVRIDIS P.: Towards automated 3d reconstruction of defective cultural heritage objects. In *Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage* (Aire-la-Ville, Switzerland, Switzerland, 2014), GCH '14, Eurographics Association, pp. 135–144. doi:10.2312/gch.20141311. 1

[HS17] HERMOZA R., SIPIRAN I.: 3d reconstruction of incomplete archaeological objects using a generative adversary network. *CoRR abs/1711.06363* (2017). arXiv:1711.06363. 1

[IHD*13] IOANNIDES M., HADJIPROCOPI A., DOULAMIS N., DOULAMIS A., PROTOPAPADAKIS E., MAKANTASIS K., SANTOS P., FELLNER D., STORK A., BALET O., JULIEN M., WEINLINGER G., JOHNSON P. S., KLEIN M., FRITSCH D.: Online 4d reconstruction using multi-images available under open access. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-5/W1* (2013), 169–174. 1

[KDD*14] KYRIAKAKI G., DOULAMIS A., DOULAMIS N., IOANNIDES M., MAKANTASIS K., PROTOPAPADAKIS E., HADJIPROCOPIS A., WENZEL K., FRITSCH D., KLEIN M.: 4d reconstruction of tangible cultural heritage objects from web-retrieved images. *International Journal of Heritage in the Digital Era 3*, 2 (2014). 1

[KL12] KERSTEN T. P., LINDSTAEDT M.: Automatic 3d object reconstruction from multiple images for architectural, cultural heritage and archaeological applications using open-source software and web services. *PFG Photogrammetrie, Fernerkundung, Geoinformation 2012*, 6 (12 2012), 727–740. doi:10.1127/1432-8364/2012/0152. 1

[LSP14] LADICKY L., SHI J., POLLEFEYS M.: Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Washington, 2014), IEEE Computer Society, pp. 89–96. 3

[LZT*11] LU M., ZHENG B., TAKAMATSU J., NISHINO K., IKEUCHI K.: Preserving the khmer smile: Classifying and restoring the faces of bayon. In *Proceedings of the 12th International Conference on Virtual Reality, Archaeology and Cultural Heritage* (2011), VAST'11, Eurographics Association, pp. 161–168. 1

[PMJh14] PARK J., MUHAMMAD T., JAE-HONG A.: The 3d reconstruction and visualization of seokguram grotto world heritage site. In *2014 International Conference on Virtual Systems Multimedia (VSMM)* (Dec 2014), pp. 180–183. doi:10.1109/VSMM.2014.7136646. 1

[SCN05] SAXENA A., CHUNG S. H., NG A. Y.: Learning depth from single monocular images. In *Advances in neural information processing systems* (Cambridge, 2005), MIT Press, pp. 1161–1168. 1

[THO*16] TANAKA S., HASEGAWA K., OKAMOTO N., UMEGAKI R., WANG S., UEMURA M., OKAMOTO A., KOYAMADA K.: See-through imaging of laser-scanned 3d cultural heritage objects based on stochastic rendering of large-scale point clouds. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III-5* (2016), 73–80. 2

[THS*12] TANAKA S., HASEGAWA K., SHIMOKUBO Y., KANEKO T., KAWAMURA T., NAKATA S., OJIMA S., SAKAMOTO N., TANAKA H., KOYAMADA K.: Particle-Based Transparent Rendering of Implicit Surfaces and its Application to Fused Visualization. In *EuroVis - Short Papers* (2012), Meyer M., Weinkaufs T., (Eds.), The Eurographics Association. 2

[TUH*14] TANAKA S., UEMURA M., HASEGAWA K., KITAGAWA T., YOSHIDA T., SUGIYAMA A., TANAKA H. T., OKAMOTO A., SAKAMOTO N., KOYAMADA K.: Application of stochastic point-based rendering to transparent visualization of large-scale laser-scanned data of 3d cultural assets. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific* (Yokohama, Japan, 2014), IEEE, pp. 267–271. 2