Feature Point Extraction and 3D Reconstruction Based on Structured-Light Pattern

You-Sin Lin, Chia-Chen Chen
National Taiwan University

Ming-Cong Su
National Taiwan University of Science and Technology

Abstract

We collect the depth data of several objects to build the dataset and use it to train our CNN-Based feature points extraction model. The result of our experiment loss can be within 1 pixel. Finally, we reconstruct 3D point clouds with triangulation on model predicted feature points.

1. Introduction

There are many ways to realize 3D sensing, and Structured-Light method is one of those. We will briefly introduce what is Structured-Light method, the pattern we use, the problem definition, and our solutions in this section.

1.1. Structured-Light Method

The sensing principle is to project the pattern on the object by projector, and capture it with camera. Then extract the feature points from the image and correspond to the feature points in the projector image. Finally, obtain the point cloud by triangulation. The following are some features of Structured-Light method.

- **Triangulation-based:** Each feature point in the images corresponds to a line in 3D space. And the point where the two rays projected by corresponding feature points intersect is the 3D coordinate of the object. Thus, we can get depth data of each pair of corresponding points in both camera and projector image plane.
- **Less dependent on scenario features than classical stereo vision-based:** Since Structured-Light method project encoded patterns on the object, it has more feature points, better robustness to ambient light than classical Stereo-Vision method.
- **High accuracy, Short working distance:** In order to clearly identify patterns, the working distance of structured light is usually short, but with self-defined patterns, this method can achieve higher depth accuracy than classical Stereo-Vision method.

1.2. Structured-Light Pattern

In Jason's paper [1], he divided the structured light pattern into two types: Time multiplexing and Spatial neighborhood.

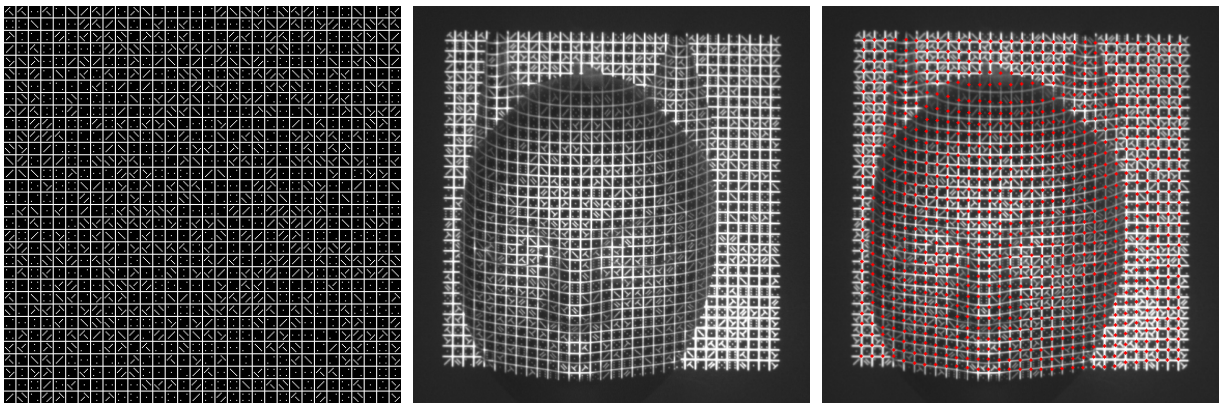
- **Time multiplexing:** Project a variety of encoded patterns in a short time, and then obtain the coordinates of the feature points on the camera plane after decoding. It is also called multiple-shot.
- **Spatial neighborhood:** Project a pseudo-randomly coded pattern so that any feature point on the pattern is unique in the camera's field of view. It is also called single-shot.

1.3. Problem Definition

We are currently using the multiple-shot pattern, but it will take a long time to collect depth data because the patterns have to be projected on the object one by one in order. And it might have multiple candidates in one feature point when using computer vision functions to extract feature points on multiple-shot patterns, which requires manual labeling.

1.4. Our Solution

First, We replaced multiple-shot patterns with single-shot pattern, and the adjacent cells of each corner are pseudo-random. Second, We trained a model to extract corner coordinates. With the model, the obtained results can be more robust and the inference time can be shorter.



(a) pattern (b) input (c) output

Figure 1. Structured light pattern / input data / result

2. Methodology

We refer to the working flow in Zhan's paper [2], and divide the task into two main processes, feature extraction and 3D reconstruction. First, features of the input image should be extracted using ResNeSt and U-Net based models. The extracted features here are the points of intersection of each corner as shown in Figure 1c. Therefore, these points can be used to reconstruct the 3D point cloud with triangulation.

2.1. Dataset

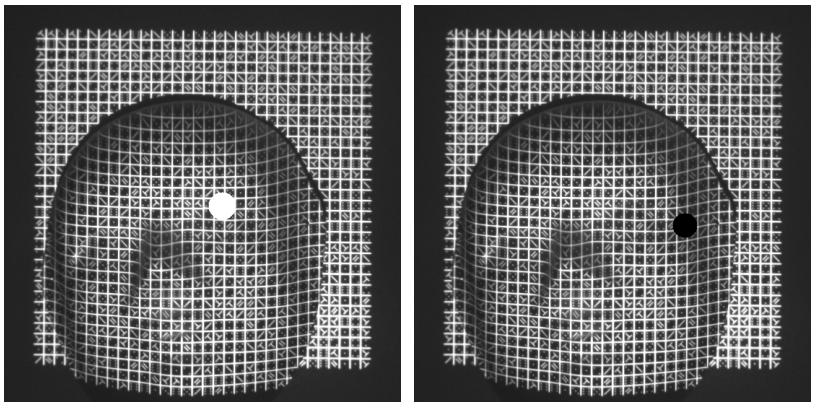
Our dataset includes 12 different smooth or sharp objects. They are respectively masks, plaster, box, and balls. We took photos of these objects at different position and rotation with the cover of structured light pattern. We also add two kinds of augmentation, reflection and occlusion, to help improve our training. Augmented data is illustrated in Figure 2. The process of generating our dataset:

1. Calibrate camera and projector
2. Project structure light pattern on the objects
3. Shoot photos with the lights off
4. Manual labeling
5. Apply augmentation

We eventually generated 210 image and label pairs divided into 7:3 for train and validation. The only testing image of our dataset is the batman mask in Figure 1b.

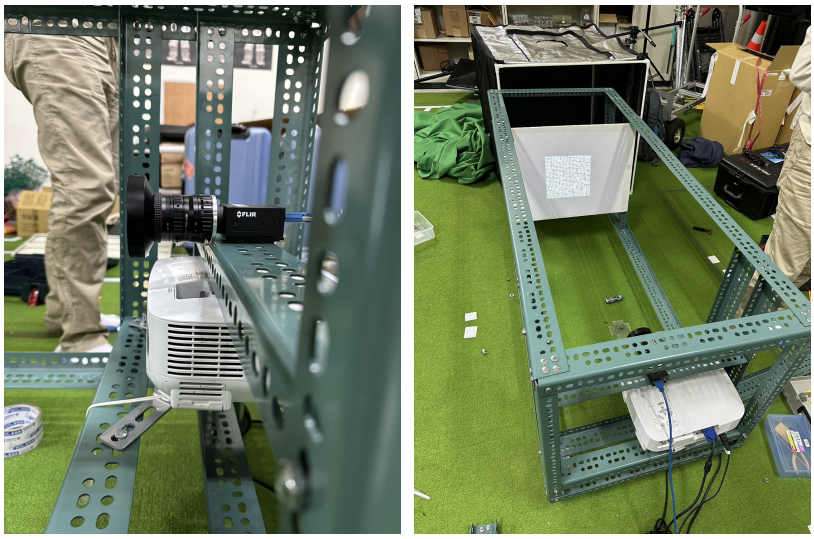
2.2. Environment Setup

We set up a camera-projector system, shown in Figure 3, to guarantee the quality of data collecting. All the components are fixed at a iron frame to ensure the consistency. Moreover, the data will be collected with the lights off to avoid environmental factors.



(a) reflection (b) occlusion

Figure 2. Augmentation



(a) camera & projector (b) whole setup

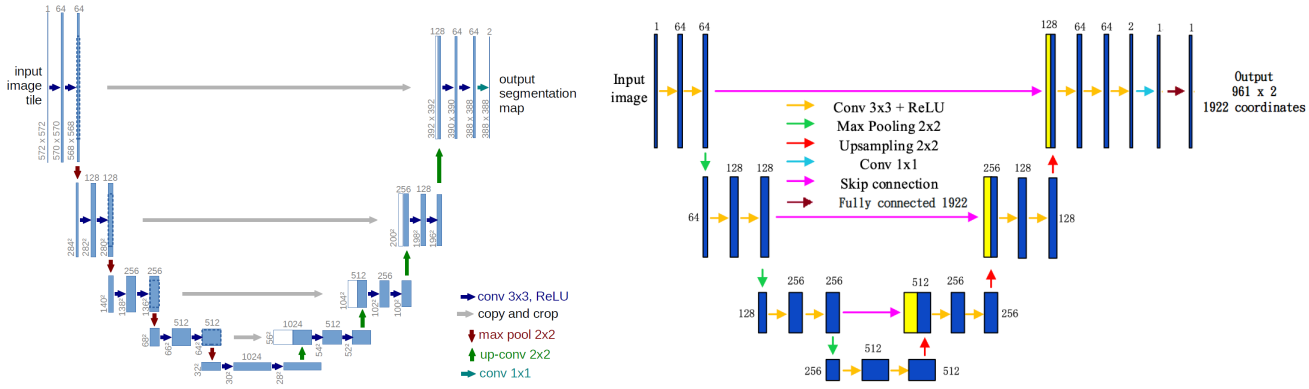
Figure 3. Environment setup

2.3. Model Architecture

Since our task need to extract feature point from 2D image first, we've searched for several CNN-based deep learning models. Finally, we use one U-Net based model and one ResNeSt model to train on our task.

- **U-Net:** In Olaf's paper [3], he proposed a CNN-based model — U-Net that performs better result on semantic segmentation tasks. The U-Net model architecture graph is shown in Figure 4a And in Dieuthuy's paper [4], they first do semantic segmentation on the image to get the grid lines and then do grid point extraction. Then, the influence of ambient light on grid point extraction can be excluded in this pipeline. There are three differences between our architecture and the original U-Net architecture:
  - **Reduced model to 3 upsamples & 3 downsamples**
  - **With padding:** To keep the size of output image the same as the input image.
  - **Add convolution layer after output segmentation map and fully-connected layer at the end of model to extract 961 points with 2D coordinates:** To extract output grid point coordinates after semantic segmentation.

  Our U-Net based model architecutre is shown in Figure 4b, with FLOPs = 167.05G and Params = 511.54M.



(a) Original U-Net architecture

(b) Our U-Net-based model architecture

Figure 4. U-Net-based model

- **ResNeSt:** The ResNet architecture is first introduced in Kaiming's paper [5]. In Hang's paper [6], he referenced ResNeXt and SKNet, integrated multi-path and feature map extraction mechanisms into the ResNeSt. ResNeSt outperforms other models on many datasets of many computer vision tasks, and we think it might perform well on feature point detection, too. Hence, we use ResNeSt to predict. ResNeSt model architecture is shown in Figure 5, with FLOPs = 103.38G and Params = 29.37M.

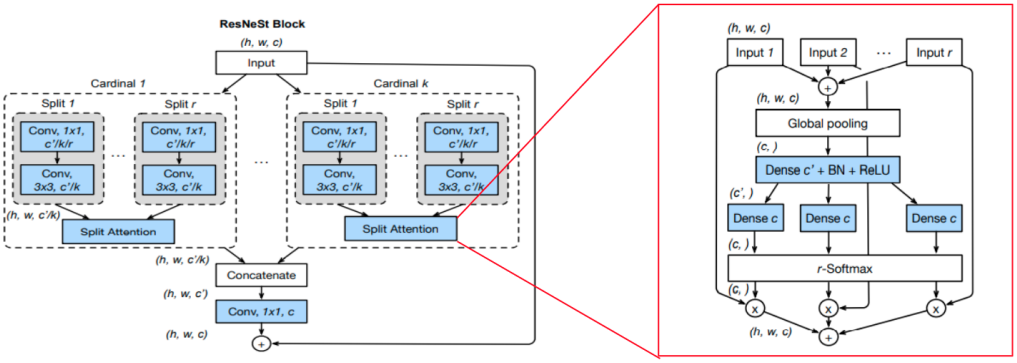


Figure 5. ResNeSt model

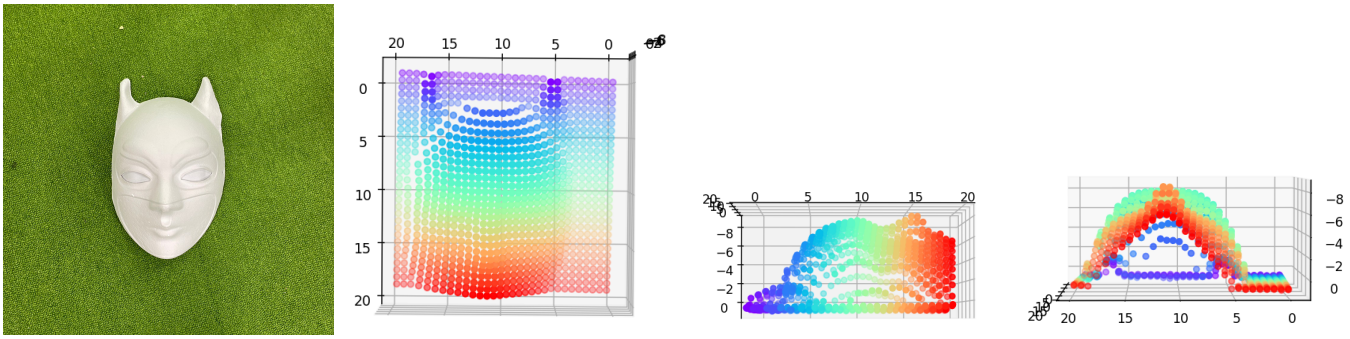
3. Experiment

- **U-Net-Based:** There are several outliers in the result of grid points extraction. And we removed these outliers according to the unexpected depth after triangulation when reconstructing the 3D point clouds. The result is shown in Figure 7.

- **ResNeSt:** The results of grid points extraction are more accurate than U-Net, but the ears of the Batman mask are not complete when the point cloud is reconstructed. The result is shown in Figure 8.

3.1. Object Reconstruction

- **Object: Bat-Man Mask**

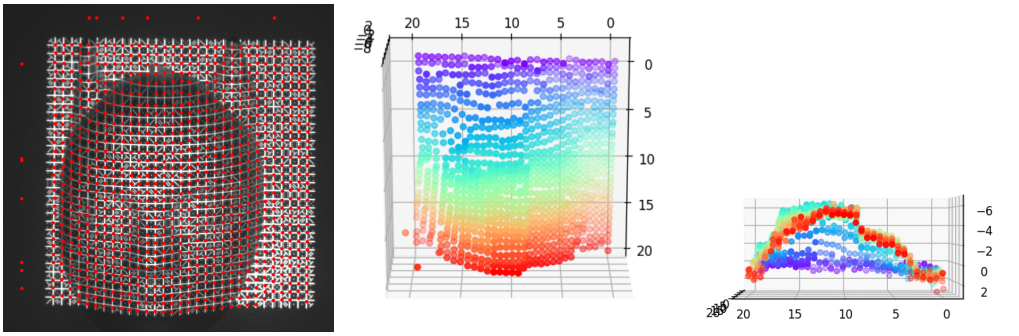


(a) Raw image (b) Front view of 3D reconstruction (c) side view of 3D reconstruction (d) bottom view of 3D reconstruction

Figure 6. Object: Bat-Man mask

– **U-Net-Based**

∗ The L2-norm loss of grid points is 4.571781337989977.

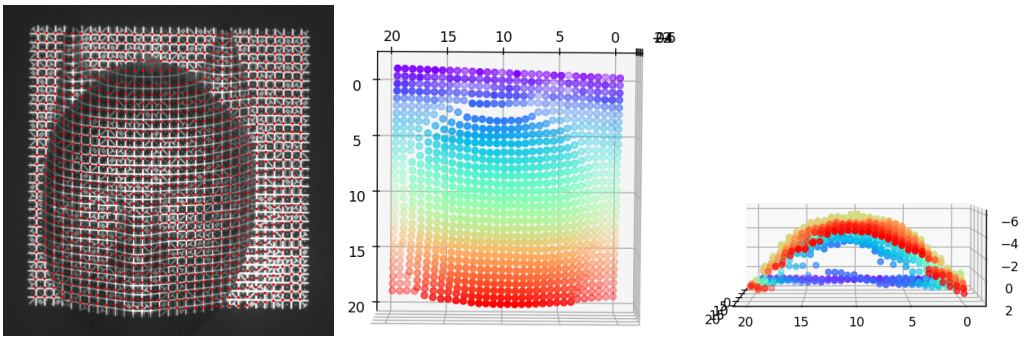


(a) Result of grid points extraction (b) Front view of 3D reconstruction (c) Bottom view of 3D reconstruction

Figure 7. U-Net-based model predict on Bat-Man mask

– **ResNeSt**

∗ The L2-norm loss of grid points is 4.405906532783523.



(a) Result of grid points extraction (b) Front view of 3D reconstruction (c) Bottom view of 3D reconstruction

Figure 8. ResNeSt based model predict on Bat-Man mask

4. Encountered Problem

- Data collecting process is complicated and the quality is unstable because of the environment.
- Triangulated 3D points can be incorrect though 2D detection is precise.
- Outliers influence the reconstruction of 3D points.
- The region whose width is only 1 to 2 grid points can't be reconstructed well(eg. The horn of the bat-man mask).

5. Discussion

We list the following reasons and possible solutions that may lead to poor results of our 3D point cloud reconstruction.

- Ambient light may influence the generation of data. At first, the undetected points and outliers are too many to label manually. After turning all the lights off, the number of detected points increase in a huge range.
- The error of feature point extraction is not small enough, leading to the wrong reconstructed point position. After the model output, we can use CV-based method to adjust the detected feature points to the sub-pixel accuracy.
- The calibration of camera and projector is not concise enough, leading to the bad reconstruction result. Before model prediction, we can evaluate our calibration result first by triangulating ground truth points to see point cloud or calculating the re-projection error.
- We can collect more diverse data including simple one, letting the model to see as much distortion as possible on every region in the structured light pattern, or we can keep fine-tune our model to better accommodate on our task.
- To improve the depth accuracy of the point cloud, we can reduce our grid size then each region may will contain more points, but it'll also make it harder for model to learn because the pixel number in each pattern will get smaller than before.

The model size of ResNeSt is far smaller than U-Net, and ResNeSt's performance on our task seems better than U-Net's. Thus, we can keep collecting data to improve ResNeSt, on the other hand, we can design a CV-based method to adjust the output of the model.

References

[1] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3:128–160, 2011. 1

[2] Zhan Song, Suming Tang, Feifei Gu, Chu Shi, and Jianyang Feng. Doe-based structured-light method for accurate 3d sensing. *Optics and Lasers in Engineering*, 2019. 2

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 3

[4] Dieuthuy Pham, Minhtuan Ha, and Changyan Xiao. A grid-point detection method based on u-net for a structured light system. *ArXiv*, abs/2012.08641, 2020. 3

[5] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 4

[6] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi-Li Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alex Smola. Resnest: Split-attention networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2735–2745, 2020. 4

Division of work

- Research : You-Sin Lin, Chia-Chen Chen, Ming-Cong Su
- Coding : You-Sin Lin, Chia-Chen Chen
- Data collecting : You-Sin Lin, Chia-Chen Chen, Ming-Cong Su
- Presentation : You-Sin Lin, Chia-Chen Chen, Ming-Cong Su