Monocular gray code 3D shape measurement based on improved Siamese Network

Yangming Li¹, *, Yan Piao¹ and Yuheng Liu¹
¹Imaging and Information Processing Laboratory, Changchun University of Science and Technology, Changchun, Jilin, 130022, China
*Corresponding author’s e-mail: 2019100483@mails.cust.edu.cn

Abstract. As 3D information is widely used in more and more industries, simplification of 3D measurement operations and higher accuracy of measurement results are important development directions in the future. Compared with the binocular measurement system, simple equipment and convenient operation are the advantages of the monocular measurement; the disadvantage is that the measurement accuracy is low and it is greatly affected by the environment. This paper proposes a method that combines deep learning with traditional monocular Gray code 3D shape measurement technology. The space transformation network is used to improve the Siamese Network, and the single-view photos are taken through spatial transformation to obtain dual-view photos, and then matched through the Siamese Network. The purpose is to use monocular measurement equipment and methods to obtain the effect of binocular measurement. The deep learning network in this article has been pre-trained with a self-made data set before use. Compared with the original method, this method has higher accuracy in plane measurement, and has a significant improvement in depth measurement, successfully reducing the depth measurement error to less than 0.1mm.

1. Introduction
The three-dimensional information of objects is an important support for human beings to understand and transform the world. Objects in the three-dimensional space have different sizes and shapes, and have different requirements for measurement technology. Among the measurement methods, the structured light method[1] is more and more widely used in modern measurement. It constitutes an important topic in computer vision and is used in many different fields.

The basic principle of the structured light technology is to project various fringe patterns on the surface of the object to be measured through a projector. The most commonly used is Gray code, because Gray code has high robustness. The deformed fringe pattern on the surface of the object to be measured is captured by the camera, and the pixels in the fringe pattern are calculated. The change of the gray value of the point can analyze the spatial position coordinates of the surface of the object to be measured, thereby obtaining the 3D data of the surface of the object to be measured.

Common high-precision three-dimensional measurement is generally binocular[2], because it can obtain more surface information of the measured object, but the disadvantages are also obvious. The binocular system must be calibrated stereoscopically before use, and captured images are twice as much as that of the monocular. Due to the need to match the feature points of the dual view, the use is relatively difficult and the running speed is affected. This paper proposes a three-dimensional shape measurement method that can obtain higher accuracy by using a monocular camera, and the projection code is Gray code. Compared with the traditional binocular Gray code three-dimensional measurement, this method
eliminates the process of binocular stereo matching, and adopts a training model based on the improved Siamese Network to achieve a virtual binocular measurement effect. Since deep learning is introduced on the basis of traditional methods, the improved Siamese Network model must be pre-trained before measurement. Finally, we combine the collected monocular data with the pre-trained network model to complete the three-dimensional surface topography measurement.

2. Structure of improved Siamese Network
This paper proposes an improved Siamese Network. Since the traditional Siamese Network needs to input matching images at both ends at the same time, this article adopts the method of monocular measurement, and there is only one image. Therefore, this article changes the monocular image spatially at the front end of the network, and uses the Spatial Transformer Network[3] proposed by Google to improve the Siamese Network. The neural network structure is shown in Figure 1. The upper part is the traditional Siamese Network[4], and the lower part is the Spatial Transformation Network. The Spatial Transformer Network transforms the pictures taken by the camera into space. The transformed picture is equivalent to another virtual camera shooting from another perspective, completing the conversion from monocular to binocular. After that, the two images are normally input into the Siamese Network. One pixel in the target Gray code image is used as a square image block with a side length of R, and D reference pixels in the reference Gray code image are used as the length R+D−1, and the width is a rectangular image block of R. In the text, the size of the target Gray code image block is 19×19, that is, R=19. According to the baseline distance between the two cameras and the statistics of the disparity data in the training set, the disparity data range of the training set objects is −50~100, so the size of the reference Gray code image block is 19×169, that is, D=151. Regarding these two image blocks as the input data of the sub-network, the basic structure of the two sub-networks consists of convolutional layers. The function of the convolutional layer is to extract the eigenvalues of the input data. Because the size of the convolution core is 3 and the padding method used is VALID, the output feature matrices after 9 convolutional layers are [1, 64] and [151,64], where 64 is the number of convolution kernels in the convolution layer, that is, the number of output features.

After the input data is processed by the weight-sharing sub-network, the feature matrix corresponding to the target pixel and the reference pixel will undergo a dot product operation to obtain the matching similarity. In deep learning, the training data needs to be labeled, because the function of the neural network structure is to find the matching point of the target pixel of the target image among the 151 reference pixels of the reference grid image. So the "label" of the neural network should be the target image that contains the matching similarity between the target pixel and the 151 reference pixels of the reference image.

3. Calibration
The measurement system usually consists of a camera, a projector, and a computer. The projector projects gray code onto the object to be measured. The camera obtains the deformed fringe pattern in
real time, and uses the known correspondence between the camera and the projector to complete it through the principle of triangulation. System calibration is to construct a known correspondence between camera, projector and space, which is an important part of structured light measurement system. The accuracy of calibration directly determines the measurement accuracy[5].

Camera calibration requires estimating the parameters of the general pinhole model which includes the intrinsic parameters, focal length, principal point, and the scale factors, as well as the extrinsic parameters defined by the rotation matrix and translation vector mapping between the world and camera coordinate systems. In particular, we use the checkerboard calibration proposed by Zhang, which is widely adopted in calibration area. Projector calibration has drawn increasing attention, in part driven by the emergence of lower-cost digital projectors. As mentioned in the front, a projector is simply the 'inverse' of a camera, so the calibration method is similar to the camera's.

4. Gray code encoding and decoding principle

Gray code patterns[6] constitute with black and white pixel values. They are also the only possibilities available with the projectors. Using such binary images requires \( \log_2(n) \) patterns to distinguish among \( n \) locations. For our projector with 1024×768 pixels, it is necessary to project the images on the scene with 10 vertical and 10 horizontal patterns, which together uniquely encode the \((u, v)\) position at each pixel.

The corresponding code value of the black stripes is 0, and the code value of the white stripes is 1, and there is only one bit difference between two adjacent code words, which has strong anti-interference ability. A set of \( M \)-bit Gray code projection patterns with a resolution of \( P \times Q \) pixels are generated as follows.

\[
g_w(i, j) = \text{fix}[2^{w-1}(j - 1)/Q + 0.5] \mod 2 \tag{1}
\]

In the above: \( g_w(i, j) \) is the gray level of the pixel \((i, j)\); \( \text{fix()} \) is the function of rounding to zero; \( w \) is the number of bits from the highest to the lowest of the Gray code, a positive integer and \( w=1, 2, 3,...,M \); \( \mod \) is the remainder symbol.

In the \( M \) corresponding Gray code images acquired by the camera, the pixel gray level \( I_w(i, j) \) corresponds to \( g_w(i, j) \). In practice, the uneven physical properties of the scene surface, irregular geometric properties, system noise in the environment and the system, etc. cause differences in the grayscale response of the pixels in the Gray code image, so it is necessary to normalize the Gray code image to reduce these effects. For this reason, add a "full dark" image and a "full bright" pattern to be projected to normalize the original Gray code image.

\[
J_w(i, j) = \frac{I_D(i, j) - I_B(i, j)}{I_B(i, j) - I_D(i, j)} \tag{2}
\]

In the above: \( I_D(i, j) \) and \( I_B(i, j) \) represent the gray scale of the "full dark" and "full bright" images respectively; \( I_w(i, j) \) represents the gray scale of the normalized Gray code image, and its range is \([0, 1]\). Then, binarize the normalized image to obtain the pixel Gray code value \( G_w(i, j) \).

\[
G_w(i, j) = \begin{cases} 1, & I_w(i, j) > 0.5 \\ 0, & I_w(i, j) \leq 0.5 \end{cases} \tag{3}
\]

Next, convert the Gray code value to the binary code value of the pixel \( B_w(i, j) \).

\[
B_w(i, j) = [\sum_{u=1}^{w} G_u(i, j)] \mod 2 \tag{4}
\]

Finally, convert the binary code to the pixel's decimal code \( k_g(i, j) \), which is the decoded value.

\[
k_g(i, j) = \sum_{w=1}^{M} B_w(i, j) \times 2^{M-w} \tag{5}
\]
From the above, according to the pixel gray scale $I_g(i,j)$ of the Gray code image, the pixel gray scale $I_R(i,j)$ and $I_w(i,j)$ in the "full dark" and "full bright" images, we can get Gray code decoded value $k_R(i,j)$ of this pixel[7].

5. Experiments
In this article, in order to improve the measurement accuracy based on the monocular camera, the training of the Siamese Network must be completed in advance before the formal measurement. Due to a lack of structured-light datasets to train the network, a dataset needs creating based on existing binocular stereo datasets. The structured-light dataset creation mainly consisted of three steps: 3D scene reconstruction, scene obtainment with random patterns, and new ground truth calculations. This article uses part of the KITTI dataset to make a small structured light training set. The data set is not easy to make, the specific process refers to SLNet proposed by Q Du[8].

| Year | Non-Occ >2 pixels | All >2 pixels | Non-Occ >3 pixels | All >3 pixels | Non-Occ >4 pixels | All >4 pixels | Non-Occ >5 pixels | All >5 pixels | End-Point | Runtime (s) |
|------|-------------------|--------------|-------------------|--------------|-------------------|--------------|-------------------|--------------|-----------|-------------|
| 2012 | 4.68              | 6.21         | 3.13              | 4.34         | 2.37              | 3.34         | 2.02              | 2.81         | 0.8px     | 1.1px       | 0.71        |
| 2015 | 9.98              | 11.69        | 7.18              | 8.92         | 5.92              | 7.66         | 4.98              | 6.68         | 1.8px     | 2.6px       | 0.35        |

After that, this article combined the traditional method to complete the following measurement. The system includes Nikon D3600 and BenQ projectors with a resolution of 1024×768. This article uses Zhang's calibration method to finish the calibration. It can be seen from the reprojection error in Figure 2 that the calibration effect of this article is better. The scattered points are roughly distributed in a circle with a small diameter.

According to the above, the stripes are projected on the surface of the object in sequence, and then the calibrated camera is used to capture the image. Part of the captured images show in Figure 3.

This paper combines deep learning with traditional methods, and adopts the improved Siamese Network's shared weight characteristics to fuse the collected monocular images into virtual binocular images with the same weight as the binocular images so as to improve the monocular structured-light accuracy of 3D topography measurement. After the image acquired after projection is converted by the

Figure 2. Reprojection error

Figure 3. Fringe projected on the object
pre-trained model, the traditional gray code three-dimensional measurement method is used to complete the final shape measurement. Finally, the depth map of the measured object and the surface model of the object are obtained in Figure 4.

![Figure 4](image_url)

Figure 4. These are the final output images of the shape measurement. The original results are on the first line and ours are on the second line.

Compared with the method before the improvement, the imaging effect near the eye socket is significantly improved. This article actually measured the maximum diameter of the orbit of the object and the maximum depth of the orbit. By comparing the three-dimensional measurement data before and after the improvement, the error between the measurement results and the actual parameters decreased from 0.14mm to 0.08mm. The measurement error is reduced to less than 0.1mm.

| Actual Value | Flat   | Depth   | Flat error | Depth error |
|--------------|--------|---------|------------|-------------|
| Original     | 3.749cm | 4.268cm | —          | —           |
| Ours         | 3.751cm | 4.260cm | 0.02mm     | 0.08mm      |

Table 2. The parameters of the measurement results

6. Conclusion and Discussion

This paper proposes a method that combines the improved Siamese Network with monocular 3D measurement, and uses a pre-trained network model to improve the measurement accuracy of monocular measurement. In order to reduce the influence of ambient light on the measurement, this paper uses the more robust Gray code fringe to test, and adds extra black and white fringe projections. The experimental results show that the three-dimensional shape measurement method proposed in this paper has higher measurement accuracy and better measurement effect in details without changing the complexity of the equipment. The pre-training of the Siamese Network uses a small data set built by ourselves, with fewer training samples, and the sample quality needs to be further improved. Considering from the image, this method still has areas worthy of improvement. On a continuously changing plane, such as a tooth, this measurement method does not get the three-dimensional information of the tooth shape. This may be the next step of the future improvement.

References

[1] Liu, Y., Fu, Y., Zhuan, Y., Zhou, P., Zhong, K., Guan, B. (2020) Large depth-of-field 3D measurement with a microscopic structured-light system. Optics Communications, Volume 481:126-145.

[2] Ye, J., Zhou, C. (2019) Time-resolved coded structured light for 3D measurement. Microwave and Optical Technology Letters, Volume 63:26-34.

[3] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K. (2016) Spatial Transformer Networks. Google DeepMind, London.

[4] Luo, W., Schwing, A., Urtasun, R. (2016) Efficient Deep Learning for Stereo Matching. In:
Computer Vision and Pattern Recognition. Las Vegas. 5695-5703.

[5] Pribanic, T., Llado, X. (2010) A state of the art in structured light patterns for surface profilometry. Pattern Recognition, Volume 43:2666-2680.

[6] Chen, B., Zhang, S. (2016) High-quality 3D shape measurement using saturated fringe patterns. Optics and Lasers in Engineering, Volume 87: 83-89.

[7] Scharstein, D., Szeliski, R. (2003) High-Accuracy Stereo Depth Maps Using Structured Light. In: Computer Vision and Pattern Recognition. Madison. 195-202.

[8] Du, Q., Liu, R., Guan, B., Pan, Y., Sun, S. (2019) Stereo-Matching Network for Structured Light. Signal Processing Letters, Volume 26: 164-168.