Single-shot dense active stereo with pixel-wise phase estimation based on grid-structure using CNN and correspondence estimation using GCN

Ryo Furukawa, Michihiro Mikamo
Hiroshima City University
Hiroshima, Japan

Ryusuke Sagawa
AIST, Japan
Tsukuba, Japan

Hiroshi Kawasaki
Kyushu University
Fukuoka, Japan

Abstract

Active stereo systems based on static pattern projection, a.k.a. oneshot scan, have been widely used for measuring dynamic scenes. Many patterns used for oneshot active stereo have grid-structures and grid-wise codes. For such systems, the grid structure is first detected, and graph matching methods are applied to estimate correspondences. However, such graph matching is often vulnerable to graph connection errors caused by grid structure analysis based on image features. Also, dense reconstruction for such systems is an open problem, where pixel-wise correspondence estimation from sparse image features is required. We propose a learning-based method to capture grid structure information and pixel-wise positional information simultaneously. We also propose to represent the grid structure by graphs with augmented connections other than 4-neighbor connections and applying them to a graph convolutional network (GCN). The proposed method can analyze large variety of grid patterns, has auto-calibration capability, can reconstruct dense shapes for fast moving objects.

1. Introduction

Among wide variety of 3D measurement methods, active stereo methods have been widely used for practical systems because of their high accuracy and fidelity. Especially, a single-frame active stereo method, which is known as oneshot scan, recently draws wide attention because it can measure moving objects, such as live humans, animals or rotating fans, or the sensor can be moved during scan, such as mounted on autonomous vehicles, air-borne drones or robotic arms. However, there are several open problems on oneshot scan, which are not solved yet. Among the problems, sparse shape reconstruction and difficulty on calibration of the system are considered critical.

Due to the sparsity of the projection pattern of oneshot scan, reconstructed shapes become inevitably sparse and an interpolation method is required. Although several efficient interpolation methods have been proposed a decade ago, there is few progress since then. In terms of calibration of oneshot scanning system, only few researches have been done, since the pattern cannot be changed and it makes the problem hard. To conduct calibration of the stereo system, correspondences between captured image and projected pattern is required. Since patterns for oneshot scan usually consist of repetitive and simple structures, such as vertical or horizontal lines for robust detection, uniqueness of local features is limited and it makes difficult to find correspondences globally.

In this paper, we propose an efficient interpolation method based on deep neural network (DNN), which can be applied to most of oneshot patterns. As mentioned above, patterns for oneshot scan usually consist of repetitive and regular structure, such as grid pattern, sinusoidal curves or dots. To interpolate spaces between those primitives, we propose a learning-based method to directly estimate "phase information" of each pixel relative to grid structure using deep neural network. One important advantage of our method is that, thanks to using deep learning method, the phase information can be extracted regardless of the actual visibility of the grid pattern. For example, in the case of line-based grid patterns, if line detection fails, the grid structure as well as phase extraction also fails in previous methods, whereas, it is possible to extract the phase information by using the surrounding information of line structures. In addition, the patterns which do not have explicit grid structure can also be interpolated by our method, which increases the generalizability of our method.

In terms of correspondence problem, a patch or a grid structure of local features have been commonly used, especially the 4-neighbor connections of the local features represented as grid graphs for oneshot scan [30, 41]. However, detection of local features is usually an unstable process, such as missing dots or broken edges because of noise, occlusion, etc. In this paper, we propose a method to make connections between adjacent local features for de-

creasing dependency to the 4-neighbor connections. The graph representing grid and code information is processed by the deep learning model of graph convolutional network (GCN) for predicting correspondences to the original pattern. Once enough number of correspondences is obtained, auto-calibration can be conducted. In summary, the contributions of the paper are as follows:

1. Learning-based grid structure detection which does not depend on the specific features of the grid pattern is proposed.
2. High-density 3D reconstruction by pixel-wise phase estimation for oneshot scan is proposed.
3. A correspondence estimation algorithm using GCN and grid-representing graph with both 4-neighbor connections and adjacency connections is proposed.

2. Related Works

2.1. Pattern coding strategy for structured light

For active stereo using a projector-camera system, pattern codification strategies can be largely classified into temporal and spatial coding approaches. Temporal encoding uses multiple images and achieve robust and pixel-wise correspondences, however, it is not suitable for capturing fast moving objects [20]. To mitigate such drawback, several techniques are proposed [19, 46, 34]. Their strategies are basically to reduce the time for capturing by changing the patterns at high speed and reducing the required number of projected patterns. Another approach is a stereo method with structured light, where multiple patterns are projected onto the object to add artificial textures on the surface of the object, achieving robust correspondence acquisition [53, 4, 49]. Recently, Gupta *et al*. proposed a method to drastically reduce the pattern under the severe condition, such as strong inter-reflection or subsurface scattering [16, 17]. However, all those methods have essential limitation in the speed of changing patterns and require special lighting devices that can change patterns at high frame rate.

Spatial encoding requires only a single image and is possible to capture fast-moving objects [30, 22, 44] and recently draw a wide attention. One severe problem for spatial encoding method is that they encode positional information into small regions, patterns tend to be complicated and easily degraded by environmental conditions, such as noise, specularity, blur, etc. Another problem is that a spatial frequency becomes inevitably low. To solve the first problem, techniques using geometric constraints have been proposed [23, 40, 36, 25, 44, 12]. Recently, deep-learning-based approaches to efficiently find correspondences for stereo pair are proposed [50, 51] and it is also applied to structured light systems. Song *et al*. proposed a method for grid detection, and a CNN is used for classifying specifically designed 256 characters embedded into the grid pattern [43]. Furukawa *et al*. used CNNs for detecting grid-structured patterns and codes embedded into the grid points of the patterns [11, 14, 13]. They also used a graph convolutional networks (GCN) [5, 31] to predict grid-wise correspondences for the detected grid [15]. For all the methods, stability of the processes has been much improved by deep-learning models, but they are all based on sparse, grid-wise reconstruction.

2.2. Dense reconstruction for oneshot scan

To deal with the sparse reconstruction problem for oneshot scan, several researchers have been conducted. Koninckx *et al*. proposed a method for dense shape reconstruction by projecting a set of stripes [24]. Methods based on periodic color codes were proposed in [21, 42, 52]. Sagawa *et al*. [39] proposed a method that projects a grid pattern and generates dense shape by interpolating correspondence between lines by Gabor filtering. All those methods basically depend on relative numbering of the dense patterns, which assumes local smoothness of the surface and linearly interpolated, and thus, it might be degraded by shape discontinuities and line detection failures. Another approach is to apply 3D smoothing filter, such as radial basis function (RBF), to point cloud to achieve dense reconstruction [33]. However, it also smooths out the high-frequency shapes of the object and create wrong shapes near occluding boundaries. In our method, we use deep neural network to recover pixel-wise phase information and achieve dense reconstruction preserving high-frequency shapes.

2.3. Auto-calibration for oneshot scan

Geometrically, optical model of video projectors and cameras are the same, *i.e*., central projection. Therefore, to calibrate projector-camera systems, typical approach is projecting temporal-coding patterns such as Gray codes onto calibration objects and applying existing camera calibration algorithms [9, 38, 54, 28, 29, 47, 45, 2, 48, 32, 1, 8, 27]. Several software packages for this approach are also available online [7, 2]. Some researchers proposed projector calibration methods in which specialized calibration patterns are unnecessary. Some use a simple white plane to calibrate multiple poses of the projector [35, 6]. In the methods mentioned above, the projected patterns are either patterns designed for calibration (*e.g*., checker grids) or temporal coding patterns such as Gray code, in order to facilitate obtaining the corresponding points. Thus, these methods cannot be used for fixed-pattern projectors designed for shape reconstruction. Furukawa *et al*. proposed a static grid pattern modulated by gap codes for oneshot scan and performed auto-calibration by imposing special pattern onto original pattern [15]. In the method, to avoid bad influences on orig-

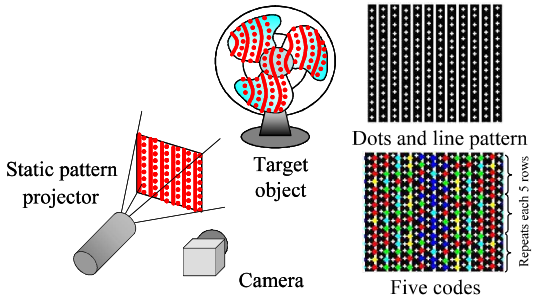
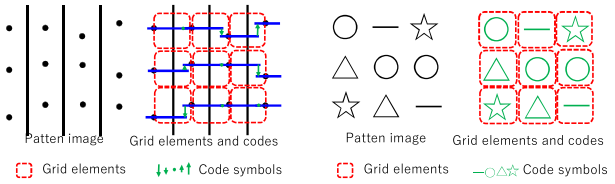
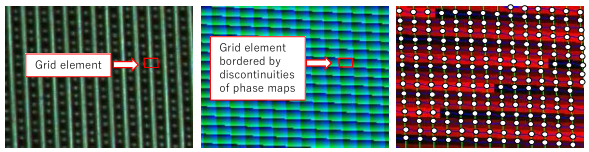


Figure 1. (Left) Scanning system, (top right) projection of patterns onto target object, (bottom right) and embedded code words.



(a) (b)

Figure 2. Examples of patterns with squared-grid structures.



(a) (b) (c)

Figure 3. An example of grid detection: (a)an input image, (b) grid features segmented by discontinuities of phase maps, and (c) graph of grid points connected by 4-neighbor connections.

inal pattern, the markers for calibration is only nine bright dots (3x3), and thus, calibration becomes unstable unless scenes are ideal condition, decreasing practicality of the system. Since there are no markers in the pattern for shape reconstruction, solution is required, which we proposed in the paper.

3. Overview

As shown in Fig.1, the proposed 3D measurement system is an active-stereo system with a camera and a static-pattern projector. The projected pattern is static and assumed to have a grid structure, where grid elements are repeated for both of the vertical and horizontal directions. Note that the grid elements are not necessarily explicit grid lines, for example, the pattern of Fig.2(a) used in [33] has vertical lines, however, it does not have horizontal lines. The pattern of Fig.2(b) used in [26] consists of several types of symbols aligned as grid structure but does not have explicit grid lines.

We also assume that each of the grid elements has a code that can be classified locally. For example, each grid element of the pattern of Fig.2(a) can be classified into 5 classes by detecting the height differences between the right and the left dots of the element. Grid elements of the pattern of Fig.2(b) can be classified by the symbols themselves.

The flow of the algorithm is shown in Fig.4. From an input image, a grid structure is detected. In this process, pixel-wise 'phase' detection and segmentation are applied. Here, phase represents angular values that represent the repetition of grids. The phase estimation is achieved by a deep learning model, *i.e.*, U-Nets, and the algorithm is applicable to wide varieties of grid patterns as shown in our experiments. Then, the grid-wise code information, *e.g.*, local dot height differences of Fig.2(a), or symbols of Fig.2(b), are also estimated by U-Nets.

The detected grid structure and codes are represented as graphs, where grid points are represented as graph nodes. Each graph node is attributed by code information and grid connections and neighbor relationships between graph nodes are represented as graph edges. The graph representing the grid structure and grid-wise code information are, then, processed by a novel deep learning model to find correspondences. For this purpose, graph convectional networks (GCN) are applied to the graph and the output of the GCN is node-wise correspondence mapping to the original patterns. Because the above correspondences are grid-wise correspondence, these correspondences are further extended to pixel-wise correspondences by using pixel-wise phase information. Finally, by using the pixel-wise correspondences, robust auto-calibration can be done and dense 3D reconstruction is achieved.

4. Region-based Grid-structure Detection

In physics and mathematics, phase $\phi(t)$ is often used to represent a periodic function, where $\phi(t)$ is a monotonic increasing angular value that can be identified with the modulus of 2π. In this paper, we use this term to represent the repetition of grid images.

To represent a 2D grid, two phase maps, $\phi_u(x, y)$ and $\phi_v(x, y)$ are used, where (x, y) is image coordinates. $\phi_u(x, y)$ represents horizontal repetition and $\phi_v(x, y)$ represents vertical repetition of the grid. Since $\phi_u(x, y)$ represents cycles, we use modulus values $\tilde{\phi_u}(x, y)$ for simplicity, where

$$\tilde{\phi_u}(x, y) \equiv \phi_u(x, y) \bmod 2\pi, \tag{1}$$

thus $\tilde{\phi_u}(x, y) \in [0, 2\pi)$. $\tilde{\phi_u}(x, y)$ is a function with sawtooth-shaped profile. Fig.3 shows examples of the source image, and its phase maps, $\tilde{\phi_u}(x, y)$ and $\tilde{\phi_v}(x, y)$. From now on, we refer $\tilde{\phi_u}(x, y)$ and $\tilde{\phi_v}(x, y)$ as phase maps.

The repetition cycles of phase maps $\tilde{\phi_u}(x, y)$ and $\tilde{\phi_u}(x, y)$ can be detected by discontinuities of the maps, where modulus value changes non-continuously from 2π to 0. Thus, the grid elements of the projected pattern can be obtained by segmentation as shown in Fig.3.

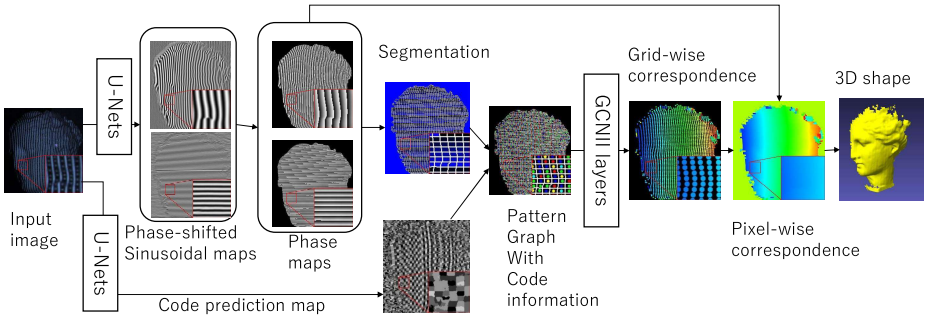


Figure 4. Dense 3D reconstruction process.

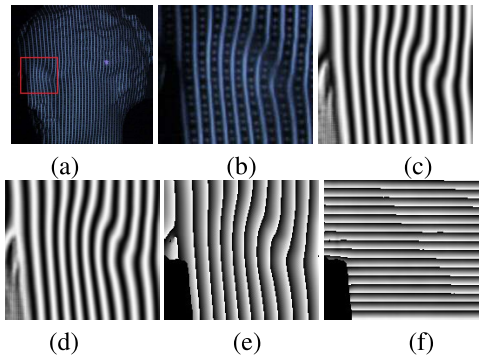


Figure 5. Estimation of phase maps using U-Nets: (a) input image with projection of Fig.2(a), (b) U-Net output for sinusoidal-pattern estimation from (b) (phase-shift:0), (c) U-Net output for sinusoidal pattern estimation from (b) (phase-shift:$\frac{4\pi}{5}$), (d) horizontal phase map from (c),(d) and other phase-shifts of $\frac{2k\pi}{5}$ for $0 \leq k \leq 4$, (e) phase map for the horizontal grid direction, (f) phase map for the vertical grid direction.

The phase maps $\tilde{\phi}_u(x, y)$ and $\tilde{\phi}_v(x, y)$ are extracted from the input image. In the previous techniques such as [39, 41], similar information is detected using Gabor filters; however, in this study, we use deep neural network, especially U-Nets [37]. The advantage of using U-Nets is that if the pattern has some repetitive structure, the U-Net can be trained to estimate phases of the repetitiveness, even if there are no explicit grid lines. For example, the pattern of Fig.2 has no horizontal lines, but the alignment of the dot features specifies the vertical cycles of a grid. Even in such cases, U-Nets can be trained to extract the cycles without requiring manual design of extraction of such features. To train the network, we generate huge amount of dataset using CG and fine-tune it, which is another contribution of the paper.

In this paper, we propose to train U-Nets to first estimate multiple 'phase-shifted sinusoidal patterns' $p_{u,k}(x, y)$ and $p_{v,k}(x, y)$. $p_{u,k}(x, y)$ is, for example, defined by

$$p_{u,k}(x, y) = (\cos(\tilde{\phi}_u(x, y) + k * 2\pi/M) + 1)/2, \quad (2)$$

where $M \geq 3$ is the number of estimated images and $0 \leq k < M$. $p_{u,k}(x, y)$ are well-known as pattern images for the phase-shift method. The advantages of estimating $p_{u,k}(x, y)$ instead of estimating $\tilde{\phi}_u(x, y)$ directly is that $p_{u,k}(x, y)$ $(0 \leq k < M)$ are continuous and smooth, whereas $\tilde{\phi}_u(x, y)$ has discontinuities. Since $p_{u,k}(x, y)$ are smooth, they and relatively easy to learn on U-Net, and that we can improve the extracted phase information using the phase shift method [18, 20], where phase information is extracted from multiple sinusoidal patterns with different sifts. An example of the output results from U-Net and an example of predicting pixel values of the phase-grid image are shown in Fig.5.

Figure 6. Samples of CG-generated training dataset: (a) 3D scene rendered with projection mapping of the dot-line pattern, (b) mask image for regions to be trained, (c) sinusoidal-pattern teacher data for horizontal phases (teacher data are provided for phase-shifts of $\frac{2k\pi}{5}$ for $0 \leq k \leq 4$, where image is a case of $k = 0$).

To estimated phase maps, $\tilde{\phi}_u(x, y)$ and $\tilde{\phi}_v(x, y)$, we apply simple segmentation algorithm so that the regions are separated by discontinuities of those maps as shown in Fig 3. The segmented regions are treated as grid elements; this means that we represent the grid as a graph, where the segmented regions are grid points, and their 4-neighbor adjacencies are represented as grid edges. If two regions have a common border where $\tilde{\phi}_u(x, y)$ is discontinuous, they are estimated to be adjacent regions for horizontal directions, and the nodes of these regions are connected by edges with left and right directions in the output graph.

For each grid point, the code information of the grid point is assigned for the correspondence estimation described later. To extract the code information from the image, a U-Net that is trained to extract code information is applied to the image, and the information vectors are sampled at the centers of the superpixels. In this paper, if each grid element has C different codes, a C-dimensional 1-hot vector is used as the feature vector of each node. $C = 5$ for

the pattern of Fig.2 (a) and $C = 4$ for Fig.2 (b).

Training of the U-Net is achieved by supervised learning. We generate training images by CG as shown in Fig.6. In the figure, (b) shows the mask region where U-Nets were trained (*i.e.*, losses on black-masked regions are ignored). Similarly, a code label image aligned with the original pattern is prepared and learned with the same approach of Nagamatsu *et al*. [33].

5. Dense 3D reconstruction with GCN-based correspondence estimation

5.1. Correspondence prediction using GCN

In the previous section, a graph representation of the detected grid is constructed, where each graph node represents a grid point of the grid and it is connected by 4-neighbor grid points. Edges for the 4-neighbor connections are attributed with four symbols of directions: UP/DOWN/LEFT/RIGHT. Each graph node is attributed by code information for each grid point. Once the graph representation of the detected grid is obtained, node-wise correspondences are predicted by using a variation of GCN.

In Furukawa *et al*. [15], GCN-based feature vector aggregation (*i.e.*, GCN layer) is done via 4-neighbor grid connections for a similar problem. However, the 4-neighbor connections often include errors by, for example, occluding boundaries, or specularities, and these errors often results in errors for correspondence prediction.

To deal with this problem, we provide the GCN model with wider information by adding new connections other than the 4-neighbor grid connections to the graph. The new connections are based on Euclidean-distance adjacencies. If a Euclidean distance between two nodes is less than a threshold, they are connected by a graph edge attributed by a symbol of ADJACENT. We call these connections as proximity connections.

Normal GCN networks are reported to have problems of over-smoothing, which make GCN with deep layers (for example, more than 5 layers) difficult to be trained properly. Chen *et al*. proposed Graph Convolutional Network via Initial residual and Identity mapping as GCNII [3]. We also use this technique for correspondence prediction for utilizing proximity connections effectively.

The l-th layer is

$$\mathbf{H}^{(l+1)} = \sigma\left(\sum_{k=1}^{5}\left((1-\alpha)\hat{\mathbf{P}}_{\mathbf{k}}\mathbf{H}^{(l)} + \alpha\mathbf{H}^{(\mathbf{0})}\right)\left((1-\beta)\mathbf{I} + \beta\mathbf{W}^{(l,k)}\right)\right) \tag{3}$$

where σ is RELU function, $k \in \{1,2,3,4,5\}$ is the index of edge labels of UP/DOWN/LEFT/RIGHT/ADJACENT, $\hat{\mathbf{P}}_{\mathbf{k}} = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{A}}_{\mathbf{k}}\hat{\mathbf{D}}^{-1/2}$ is the graph convolution matrix calculated from adjacency matrix for each edge labels $\hat{\mathbf{A}}_{\mathbf{k}}$, and α and β are hyper-parameters for GCNII. We set $\alpha = 0.5$ and $\beta = 0$ based on our experiments. The initial node feature $\mathbf{H}^{(0)}$ is given to each of the stacked layers as a skip connection.

At the output of the stacked layers, feature vectors are processed by dense linear layer and transformed to have P channels, where the candidates of correspondence ID are integers from 0 to $P-1$. P is the number of grid points in the original pattern grid. For the pattern shown in Fig.1 and Fig.2(a), each detected node is classified into 5×120 classes, *i.e.*, the pattern has 5 repetitions of grid in the vertical direction, and 120 independent features in the horizontal direction.

The training of GCN is done with supervised manner. The training dataset is generated by applying the U-Nets explained in Sec. 4 to CG-generated images shown in Fig.6, and making graph representation of the pattern by segmenting the phase images.

5.2. Auto-calibration and Dense 3D reconstruction

A node of the constructed graph represents a detected feature of grid point. After node-wise correspondence estimation of the graph, each of the features is mapped to a grid point on the original pattern. Also, the pixel-wise phase values stored within the grid region are positional information within the corresponding grid point.

By integrating this information, pixel-wise correspondences to the original pattern can be estimated, and, consequently, dense 3D depth is obtained. This process is the same is 'phase unwrapping' of the phase information obtained in Fig.5 using global positional information predicted by the GCN. With the phase unwrapping, we can estimate dense maps representing x-coordinates of the projector for each camera pixel. Then, we can obtain dense 3D reconstruction by using the light sectioning method.

In situations where the parameters of the projector-camera systems are unknown, we should calibrate the projector-camera system. By using GCN-based correspondence estimation, calibration is possible, since GCN enables estimation of 2D correspondences without using epipolar constraints (auto-calibration). We applied RANSAC for auto-calibrating the extrinsic parameters. First, 8 correspondences are sampled from the correspondence predictions, and the epipolar constraint errors are estimated for the rest of the correspondence predictions. If the ratio of outliers is below a pre-defined threshold, the epipolar constraint errors are minimized.

Once the extrinsic parameters are obtained, we can limit the correspondence candidates around the epipolar lines, as shown in Fig.8. In the process of applying the epipolar constraints, a grid point from the image (*e.g.*, the sample point in Fig.8(left)), has corresponding epipolar line in the pattern image. Blue points in Fig.8(right) are grid points of the

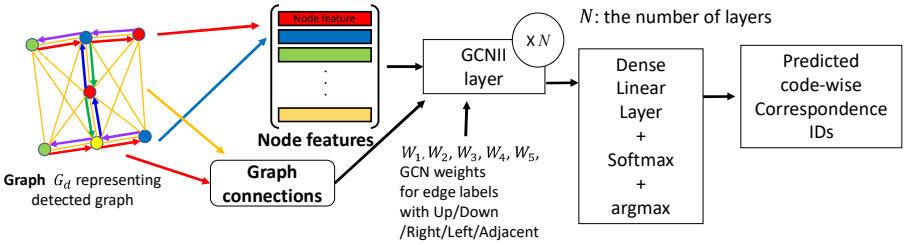


Figure 7. Node-to-node similarity calculation between detected graphs and the graph of the projected pattern.

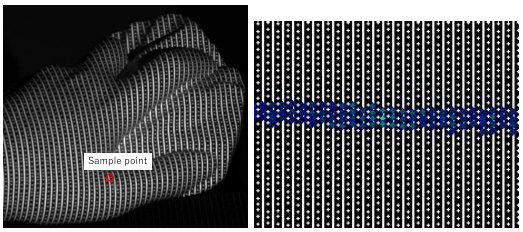


Figure 8. Applying epipolar constraints. The left image is a captured image. Point A in the left image can be limited by using epipolar constraints. The right image is projected pattern limited by epipolar constraints using estimated extrinsic parameters. The corresponding point can be found from the limited candidates colored by blue in the right image. And the corresponding point can be found from the limited candidate marked by cyan in the right.

Table 1. RMSE between the ground truth and inferences. The unit of the errors is projector pixels.

	Direct	2 phases	5 phases
Profile of A	0.719	0.215	0.190
Profile of B	0.755	0.450	0.199

pattern near the epipolar line. Then, the correspondence of the sample point is first limited to the blue points and then taking argmax of the GCN output in the limited candidate set.

6. Experiment results

6.1. Evaluation of phase estimation

For phase detection, we proposed to predict phase-shifted sinusoidal patterns $p_{u,k}(x, y)$ by U-Nets (Eq. 2) for accurate phase calculation. To show the effectiveness of the method, we estimate horizontal phase maps $\tilde{\phi}_u(x, y)$ for the image of Fig.9(a) by three methods. The ground truth of the phase is obtained by applying phase shift pattern projection and is shown in Fig.9(b). The baseline method shown in Fig.9(c) is simply to train a U-Net to output sawtooth-shaped $\tilde{\phi}_u(x, y)$. The second method shown in Fig.9(d) is estimating $\tilde{\phi}_u(x, y)$ with two sinusoidal patterns, and the third in Fig.9(e) is estimating $\tilde{\phi}_u(x, y)$ with five sinusoidal patterns. Fig.9(f) and (e) shows the profile values at the lines shown in Fig.9(a). For (f) and (e), the left column

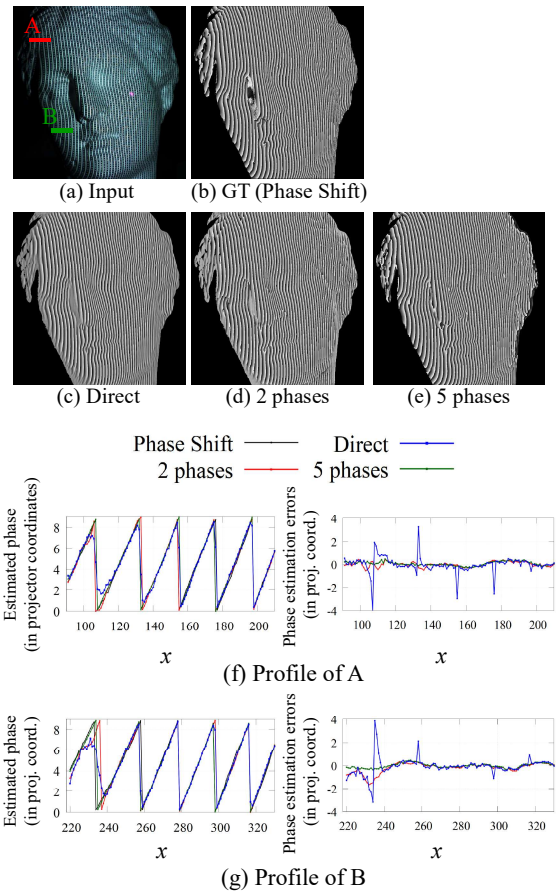


Figure 9. Evaluation of phase accuracies obtained by U-Nets.

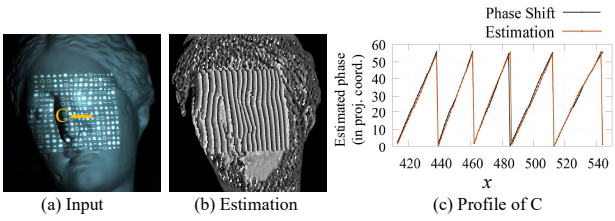


Figure 10. Phase estimation for grid pattern with Fig.2(b).

shows the profiles and the right column shows the errors from the ground-truth values measured by the phase-shift method.

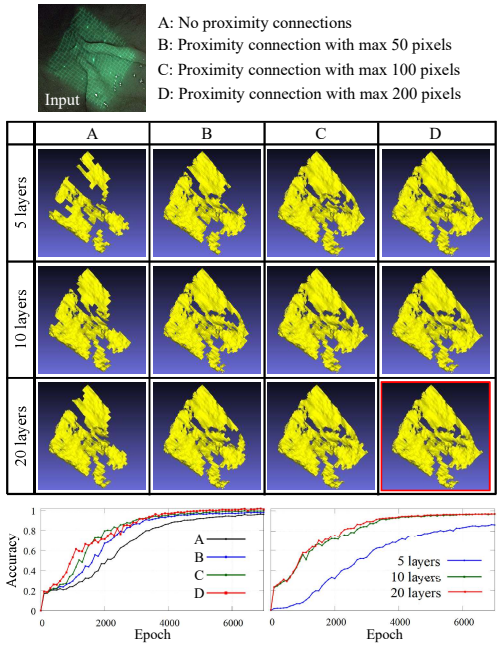


Figure 11. Effects of the number of GCN layers and proximity connections.

The Results of direct estimation are shown to have large errors at the discontinuities of $\tilde{\phi}_u(x, y)$. For the comparison of 2-phase and 5-phase method, the 5-phase method produced in more stable results than 2-phase method near the rim of the projected pattern. At those regions, images generated by the U-Nets may include artifacts or dimmer results. For the 5-phase method, those effects were corrected more efficiently than the 2-phase method.

The RMSEs of the estimated phases compared to ground truth are shown in Tab.1. The result shows the effectiveness of using 5 images for phase detection.

Fig.10 shows an example of phase detection for the pattern shown in Fig.2(b). It shows that the proposed method can detect phases with arbitrary patterns with grid structures.

6.2. Evaluation of GCN-based correspondence prediction

To show the effects of the number of GCN layers and proximity connections explained in Sec. 5.1, a pattern-projected endoscopic image, shown in the top-left cell of Fig.11, is reconstructed for various conditions, where GCNII layers are stacked with 5/10/20 layers, with and without proximity connections, and with max distances of proximity connections of 50/100/200 pixels. Then, erroneously reconstructed regions are manually deleted and shown in the table. Thus, in the cells of Fig.11, the more regions are shown, the better the results. Those results show that the models with deeper layers showed better performance,

Table 2. Comparison of the extrinsic-parameter auto-calibration results of the proposed method and the pattern [10], where 9 detectable bright markers are placed in the projected pattern. Planes with the right angle are measured and the estimated angles are shown.

	Proposed (2 images)	9 markers[10] (3 images)	9 markers[10] (9 images)
Angles	90.84	116.47	77.89

Also, the accuracy changes in the training processes are shown in for different Fig.11, with cases of the numbers of layers (5/10/20) and different proximity connection cases.

The results show that proximity connections increased the accuracies, although differences between proximity connections with max 100 pixels and with max 200 pixels were not apparent. Generally, more layers and more proximity connections improved the accuracies.

6.3. Auto-calibration of projector-camera system

By using the GCN, we can predict correspondences of an uncalibrated projector-camera system effectively. This can be used for on-the-fly calibrations. To demonstrate this, we capture data using a projector-camera system using a video projector that can project arbitrary images, and calibrated the projector-camera extrinsic parameters using the method proposed in Furukawa *et al.* [10], where 9 detectable bright markers are placed in the pattern. Then, we projected the proposed pattern and auto-calibrated the projector-camera system using the correspondences obtained from the GCN. To evaluate the obtained parameters, we measured a cube-shaped object with the right angle and evaluated the obtained angles using plane-fitting. Tab. 2 shows that the proposed method obtained the parameters better than Furukawa *et al.* [10] with fewer image captures. Note that we use 2 images just for increasing the number of correspondences, and a single-image auto-calibration is also possible.

6.4. Comparison to previous techniques

To evaluate the dense 3D reconstruction, the results of the proposed method is compared to other methods of waved-grid pattern from Sagawa *et al.* [41], and phase-shifting method[20] using multiple-capture as the Ground Truth shape. Fig.12 shows the reconstruction results. For the method of waved-grid pattern pixel-wise image matching is used for dense reconstruction. The quality of the proposed method was better than the waved-grid.

Fig.13 shows an example of measuring a fast-moving object (a rotating fan). The result is compared with measurement of Kinect v2 sensor. Using a fast shutter speed, the proposed method could capture the shape of the fan that is moving very fast, whereas there are artifacts for the result of Kinect v2 caused by the motion.

The proposed method does not necessarily rely on epipolar constraints. Thus it is more insensitive to calibration er-

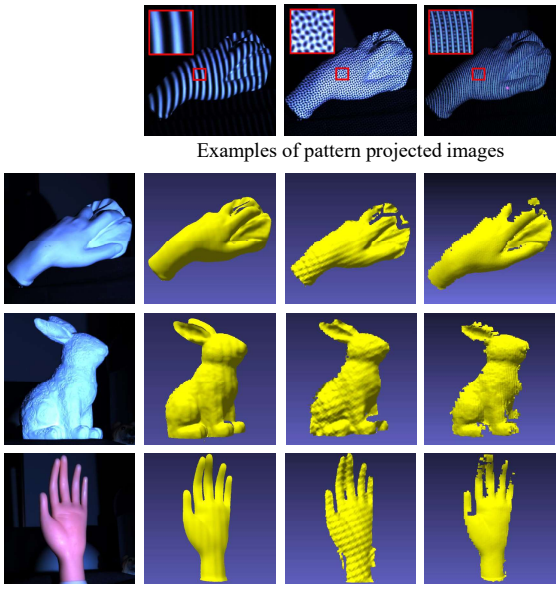


Figure 12. Dense 3D reconstruction results by using a video projector: (first column) examples of pattern projected images. (a): object appearances, (b): result of phase-shift method (GT), (c): results of Sagawa *et al.* [41], and (d): proposed.

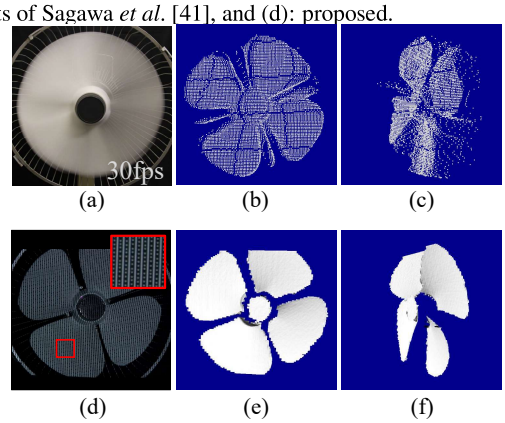


Figure 13. Measurement of fast moving object (a rotating fan). (a): Image captured by normal shutter speed. (b)(c): Measurement results by Kinect v2. (d): Image captured by fast shutter speed. (e)(f): Measurement results by the proposed method.

rors. Fig.14 and Fig.15 shows that the proposed method is insensitive to calibration errors of the projector camera system, whereas Sagawa *et al.* [41] could not reconstruct the shape with calibration errors.

Fig.16 shows a reconstruction example of a textured object. The proposed method successfully reconstructed textured dolls.

7. Conclusion

We proposed a method for 3D reconstruction for a single-shot active-stereo system, where a static pattern with a grid structure and grid-wise code information is projected. From the captured image, the grid structure is detected as phase maps by U-Nets. The phase maps are used both for grid structure analysis and pixel-wise positional information in grid elements, achieving dense 3D reconstruction. Also, for grid-wise correspondence estimation, grid structure and codes were represented as graphs with 4-neighbor connections and proximity connections, to which deep learning models of GCNII layers were applied. Experiments show that the adjacency connections and deep-layered GCNII were effective to increase the stability of reconstruction and accurate dense 3D shapes was achieved. Realtime system as well as efficient noise removal are our future work.

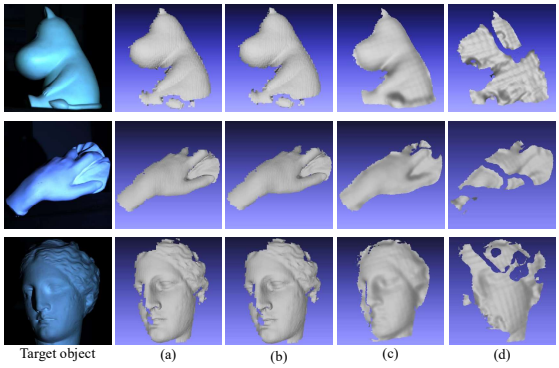


Figure 14. The reconstruction results by the proposed method and the method by the waved-grid pattern projection [41]. From left to right, the target object is reconstructed by (a) the proposed method with the original parameters, (b) that with the parameters including noise, (c) the method by wave-pattern projection with the original parameters, and (d) that with the parameters including noise.

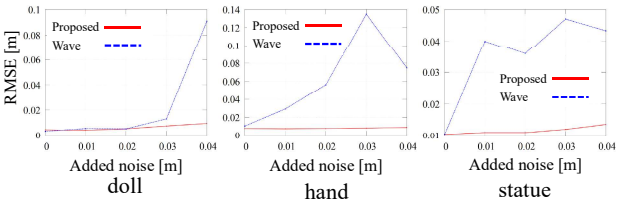


Figure 15. RMSE of Fig.14. It is clearly shown that when noises are given to the projector camera poses, the RMSE is rapidly getting worse on conventional methods, whereas same RMSEs are kept on our GCN based method.

Figure 16. A reconstruction example of a textured object.

Acknowledgment

This work was supported by JSPS/KAKENHI 20H00611, 18K19824, 18H04119, and NEDO(JPNP20006) in Japan.

References

[1] Andrea Albarelli, Luca Cosmo, Filippo Bergamasco, and Andrea Torsello. High-coverage 3d scanning through online structured light calibration. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 4080–4085. IEEE, 2014.

[2] Samuel Audet and Masatoshi Okutomi. A user-friendly method to geometrically calibrate projector-camera systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 47–54. IEEE, 2009.

[3] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning*, pages 1725–1735. PMLR, 2020.

[4] J DAVIS. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(2):296–302, 2005.

[5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[6] Jamil Draréni, Sébastien Roy, and Peter Sturm. Geometric video projector auto-calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 39–46. IEEE, 2009.

[7] G Falcao, Natalia Hurtos, J Massich, and D Fofi. Projector-camera calibration toolbox. *Erasumus Mundus Masters in Vision and Robotics*, 2009.

[8] Oliver Fleischmann and Reinhard Koch. Fast projector-camera calibration for interactive projection mapping. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 3798–3803. IEEE, 2016.

[9] Ryo Furukawa and Hiroshi Kawasaki. Uncalibrated multiple image stereo system with arbitrarily movable camera and projector for wide range scanning. In *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pages 302–309. IEEE, 2005.

[10] Ryo Furukawa, Ryunosuke Masutani, Daisuke Miyazaki, Masashi Baba, Shinsaku Hiura, Marco Visentini-Scarzanella, Hiroki Morinaga, Hiroshi Kawasaki, and Ryusuke Sagawa. 2-dof auto-calibration for a 3d endoscope system based on active stereo. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7937–7941. IEEE, 2015.

[11] Ryo Furukawa, Daisuke Miyazaki, Masashi Baba, Shinsaku Hiura, and Hiroshi Kawasaki. Robust structured light system against subsurface scattering effects achieved by cnn-based pattern detection and decoding algorithm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[12] Ryo Furukawa, Hiroki Morinaga, Yoji Sanomura, Shinji Tanaka, Shigeto Yoshida, and Hiroshi Kawasaki. Shape acquisition and registration for 3d endoscope based on grid pattern projection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2016.

[13] Ryo Furukawa, Genki Nagamatsu, and Hiroshi Kawasaki. Simultaneous shape registration and active stereo shape reconstruction using modified bundle adjustment. In *2019 International Conference on 3D Vision (3DV)*, pages 453–462. IEEE, 2019.

[14] R. Furukawa, G. Nagamatsu, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, and H. Kawasaki. Simultaneous shape and camera-projector parameter estimation for 3d endoscopic system using cnn-based grid-oneshot scan. *Healthcare Technology Letters*, 6(6):249–254, 2019.

[15] Ryo Furukawa, Shiro Oka, Takahiro Kotachi, Yuki Okamoto, Shinji Tanaka, Ryusuke Sagawa, and Hiroshi Kawasaki. Fully auto-calibrated active-stereo-based 3d endoscopic system using correspondence estimation with graph convolutional network. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4357–4360. IEEE, 2020.

[16] Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, and Srinivasa G Narasimhan. Structured light 3d scanning in the presence of global illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 713–720. IEEE, 2011.

[17] Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, and Srinivasa G Narasimhan. A practical approach to 3d scanning in the presence of interreflections, subsurface scattering and defocus. *International Journal of Computer Vision (IJCV)*, 102(1-3):33–55, 2013.

[18] Erwin Hack and Jan Burke. Invited review article: measurement uncertainty of linear phase-stepping algorithms. *Review of Scientific Instruments*, 82(6):061101, 2011.

[19] Olaf Hall-Holt and Szymon Rusinkiewicz. Stripe boundary codes for real-time structured-light range scanning of moving objects. In *Proceedings of the International Conference on Computer Vision (ICCV)*, volume 2, pages 359–366. IEEE, 2001.

[20] Katsushi Ikeuchi, Yasuyuki Matsushita, Ryusuke Sagawa, Hiroshi Kawasaki, Yasuhiro Mukaigawa, Ryo Furukawa, and Daisuke Miyazaki. Active lighting and its application for computer vision.

[21] Changsoo Je, Sang Wook Lee, and Rae-Hong Park. High-contrast color-stripe pattern for rapid structured-light range imaging. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 95–107, 2004.

[22] Hiroshi Kawasaki, Ryo Furukawa, Ryusuke Sagawa, and Yasushi Yagi. Dynamic scene shape reconstruction using a single structured light pattern. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. Ieee, 2008.

[23] Hiroshi Kawasaki, Satoshi Ono, Yuki Horita, Yuki Shiba, Ryo Furukawa, and Shinsaku Hiura. Active one-shot scan for wide depth range using a light field projector based on coded aperture. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3568–3576, 2015.

[24] Thomas P. Koninckx and Luc Van Gool. Real-time range acquisition by adaptive structured light. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(3):432–445, March 2006.

[25] Thomas P Koninckx and Luc Van Gool. Real-time range acquisition by adaptive structured light. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(3):432–445, 2006.

[26] Yang Lei, Kurt R Bengtson, Lisa Li, and Jan P Allebach. Design and decoding of an m-array pattern for low-cost structured light 3d reconstruction systems. In *2013 IEEE International Conference on Image Processing*, pages 2168–2172. IEEE, 2013.

[27] Chunyu Li, Yusuke Monno, Hironori Hidaka, and Masatoshi Okutomi. Pro-cam ssfm: Projector-camera system for structure and spectral reflectance from motion. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2414–2423, 2019.

[28] Zhongwei Li, Yusheng Shi, Congjun Wang, and Yuanyuan Wang. Accurate calibration method for a structured light system. *Optical Engineering*, 47(5):053604, 2008.

[29] Jiarui Liao and Lilong Cai. A calibration method for uncoupling projector and camera of a structured light system. In *2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 770–774. IEEE, 2008.

[30] Microsoft. Xbox 360 Kinect, 2010. http://www.xbox.com/en-US/kinect.

[31] Michihiro Mikamo, Hiroshi Kawasaki, Ryusuke Sagawa, and Ryo Furukawa. Gcn-calculated graph-feature embedding for 3d endoscopic system based on active stereo. In *International Workshop on Frontiers of Computer Vision*, pages 253–266. Springer, 2021.

[32] Daniel Moreno and Gabriel Taubin. Simple, accurate, and robust projector-camera calibration. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 464–471. IEEE, 2012.

[33] Genki Nagamatsu, Ryo Furukawa, Ryusuke Sagawa, and Hiroshi Kawasaki. Single-wavelength and multi-parallel dotted-and solid-lines for dense and robust active 3d reconstruction. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE, 2019.

[34] Srinivasa G Narasimhan, Sanjeev J Koppal, and Shuntaro Yamazaki. Temporal dithering of illumination for fast active vision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 830–844. Springer, 2008.

[35] Takayuki Okatani and Koichiro Deguchi. Autocalibration of a projector-camera system. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(12):1845–1855, 2005.

[36] Marc Proesmans and Luc Van Gool. One-shot 3d-shape and texture acquisition of facial data. In *Audio-and Video-based Biometric Person Authentication*, pages 411–418. Springer, 1997.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[38] Filip Sadlo, Tim Weyrich, Ronald Peikert, and Markus Gross. A practical structured light acquisition system for point-based geometry and texture. In *Proceedings Eurographics/IEEE VGTC Symposium Point-Based Graphics, 2005.*, pages 89–145. IEEE, 2005.

[39] Ryusuke Sagawa, Hiroshi Kawasaki, Shota Kiyota, and Ryo Furukawa. Dense one-shot 3d reconstruction by detecting continuous regions with parallel line projection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1911–1918. IEEE, 2011.

[40] Ryusuke Sagawa, Yuichi Ota, Yasushi Yagi, Ryo Furukawa, Naoki Asada, and Hiroshi Kawasaki. Dense 3d reconstruction method using a single pattern for fast moving object. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1779–1786. IEEE, 2009.

[41] Ryusuke Sagawa, Kazuhiro Sakashita, Nozomu Kasuya, Hiroshi Kawasaki, Ryo Furukawa, and Yasushi Yagi. Grid-based active stereo with single-colored wave pattern for dense one-shot 3d scan. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 363–370. IEEE, 2012.

[42] Joaquim Salvi, Joan Batlle, and El Mustapha Mouaddib. A robust-coded pattern projection for dynamic 3D scene measurement. *Pattern Recognition*, 19(11):1055–1065, 1998.

[43] Lifang Song, Suming Tang, and Zhan Song. A robust structured light pattern decoding method for single-shot 3d reconstruction. In *2017 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 668–672. IEEE, 2017.

[44] Ali Osman Ulusoy, Fatih Calakli, and Gabriel Taubin. One-shot scanning using de bruijn spaced grids. In *Proceedings of the International Conference on Computer Vision (ICCV) Workshops*, pages 1786–1792. IEEE, 2009.

[45] GAO Wei, WANG Liang, and HU Zhan-Yi. Flexible calibration of a portable structured light system through surface plane. *Acta Automatica Sinica*, 34(11):1358–1362, 2008.

[46] Thibaut Weise, Bastian Leibe, and Luc Van Gool. Fast 3d scanning with automatic motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[47] Koichiro Yamauchi, Hideo Saito, and Yukio Sato. Calibration of a structured light system by observing planar object from unknown viewpoints. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1–4. IEEE, 2008.

[48] Shuntaro Yamazaki, Masaaki Mochimaru, and Takeo Kanade. Simultaneous self-calibration of a projector and a camera using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 60–67. IEEE, 2011.

[49] Mark Young, Erik Beeson, James Davis, Szymon Rusinkiewicz, and Ravi Ramamoorthi. Viewpoint-coded structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007.

[50] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015.

[51] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches.

The journal of machine learning research, 17(1):2287–2318, 2016.

[52] L. Zhang, B. Curless, and S. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *3DPVT*, pages 24–36, 2002.

[53] Li Zhang, Noah Snavely, Brian Curless, and Steven M Seitz. Spacetime faces: high resolution capture for modeling and animation. In *Proceedings of SIGGRAPH*, pages 548–558. 2004.

[54] Song Zhang and Peisen S Huang. Novel method for structured light system calibration. *Optical Engineering*, 45(8):083601, 2006.