2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Automatic Feature Extraction using CNN for Robust Active One-shot Scanning

Ryusuke Sagawa
Intelligent Systems Research Institute,
National Institute of Advanced
Industrial Science and Technology
Ibaraki, Japan

Yuki Shiba, Takuto Hirukawa,
Satoshi Ono, Hiroshi Kawasaki
Information and Biomedical Engineering,
Kagoshima University
Kagoshima, Japan

Ryo Furukawa
The Graduate School of
Information Science,
Hiroshima City University
Hiroshima, Japan

***Abstract*—Active one-shot scanning techniques have been widely used for various applications. Stereo-based active one-shot scanning embeds a positional information regarding the image plane of a projector onto a projected pattern to retrieve correspondences entirely from a captured image. Many combinations of patterns and decoding algorithms for active one-shot scanning have been proposed. If the capturing environment lacks the assumed conditions, such as the absence of strong external lights, then reconstruction using those methods is degraded, because the pattern decoding fails. In this paper, we propose a general reconstruction algorithm that can be used for any kind of patterns without strict assumptions. The technique is based on an efficient feature extraction function that can drastically reduce redundant information from the raw pixel values of patches of captured images. Shapes are reconstructed by efficiently finding correspondences between a captured image and the pattern using low-dimensional feature vectors. Such a function is created automatically by a convolutional neural network using a large database of pattern images that are efficiently synthesized by using GPU with wide variation of depth and surface orientation. Experimental results show that our technique can be used for several existing patterns without any ad hoc algorithm or information regarding the scene or the sensor.**

I. Introduction

Active 3D measurement techniques have been studied for a long time. Techniques for acquiring the 3D shapes of dynamic objects from dynamically moving sensors becomes important for various critical applications, such as self-driving cars, motion analysis, medical diagnosis, and virtual reality systems. One-shot active 3D scanning techniques have been widely studied, and actual systems have been built and used in real applications [1], [2], [3]. From a technical point of view, those systems can be categorized into two types, active stereo systems and time of flight (TOF) systems. Although both techniques have advantages and disadvantages, we propose in this paper a new active stereo system due to the advantages on such systems have over TOF systems in regard to energy efficiency, high resolution, and stable reconstruction.

Active stereo systems are based on encoding positional information of the image plane of the projector onto its projected pattern, after which 3D shapes are reconstructed by decoding them from the captured image [4]. Two types of approaches are known for this encoding/decoding process: spatial and temporal. Since temporal methods require multiple frames, they are not suitable for one-shot scanning, and thus, spatial methods, which encode the information into a specific area of the pattern and require just a single frame, have been intensively studied and commercialized.

One problem with active stereo methods based on spatial encoding is difficulty in deciding the best set of pattern and decoding algorithms for various environments, such as the baseline between pattern projector and camera, lighting conditions, the texture, material, and shape frequency of the surfaces of target objects, as well as the depth range of the projector and camera. Since a general solution for such large variations does not yet exist, each technique usually assumes its own conditions for its own purpose, and there are no standard and clear criteria for evaluating the appropriate patterns and algorithms for specific conditions. This creates difficulty even for experts in finding the best set of patterns and algorithms for the specific purposes. Our purpose is to provide a general algorithm not dependent on such conditions, to allow users to concentrate on finding the best patterns for their purposes.

The decoding of a system can be regarded as the extraction of a necessary, compact, and low-dimensional representation of positional information from high-dimensional raw input signals with strong redundancy. Universal algorithms for such problems have been intensively studied and formulated in other areas, such as file compression and pattern recognition. Our method follows the same approach for extracting 2D positional information from small patches of captured images by using machine learning techniques. Unlike previous techniques, which usually assume special noise or distortion models depending on their purposes, such complicated assumptions are avoided by our technique by just preparing exemplars for learning. In the experiments, we show that our method is comparable to several one-shot algorithms without any manual interventions or information regarding the sensor and its settings.

II. Related work

3D surface reconstruction methods based on active stereo using coded structured light have been widely studied in terms of their practical usefulness. These methods are typically categorized as temporal or spatial coding techniques [4].

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

Temporal coding methods are superior to spatial resolutions, but inferior to temporal resolutions.

Research into spatial coding has increased in recent years, because of growing demands for dynamic scene capturing. Some methods use color coding [5], [6], [7], [8]. One problem of color codifications is that the results are often affected by the surface colors and/or textures. Some researchers use PCA for the color space analysis [6], [9] or clustering approaches [8] to cope with this problem. As another approach, dependencies on color information can be reduced by using geometrical characteristics of structured patterns [1], [10], [11]. The geometrical characteristics of patterns include random-dot [1], or grid-structured patterns [10], [11].

Most previously active stereo methods have been based on explicitly coded patterns, in which the positional information has been embedded into the pattern images using rules. However, with increases in the computational power of PCs, matching-based active stereo methods, in which captured images and pattern signals are directly compared and matched, have been studied recently [12], [13], [14]. These methods can be considered as "implicit" coding of projected patterns.

For those example- or matching-based approaches, comparing image patches allowing variations caused by depth and normal changes is important. Such a problem is one of the main topics of computer vision research and normalized cross correlation (NCC) is a standard technique for passive stereo. Another widely used solution is image matching in the eigenspace of the set of sample images with those variations. Such approaches have been used for face recognition with various changes in illumination or facial expressions [15], [16], object recognition with various pose changes in 3D space [17], and fast image matching using dimensionality reduction [18].

Some recent new approaches for image matching allowing variations in appearances are learning based. Convolutional neural networks (CNN) [19] have been proven to provide a powerful method for image recognition tasks. While eigenspace-based methods are constrained to linear transformations for dimensionality reduction, CNNs can be trained to estimate low-dimensional representations of model images that use nonlinear transformations of image patches. Some recent research uses CNNs for comparing image patch pairs of stereo images for dense estimations of depth images [20], [21], but not in active stereo.

III. Learning-based pattern encoding/decoding

A. System configuration

We assume a system that consists of a single pattern projector and camera, as shown in Fig. 1(a). The camera and the projector are assumed to be calibrated (*i.e.*, the intrinsic parameters of the devices and their relative positions and orientations are known). Since the projector casts a static pattern, no synchronization is required, and it is suitable for acquiring a 3D shape of a dynamic scene. In terms of the projecting pattern, we assume only a single color, which has advantages for the simple construction of a pattern projector and robust detection of a pattern over texture as shown in Fig. 1(b).

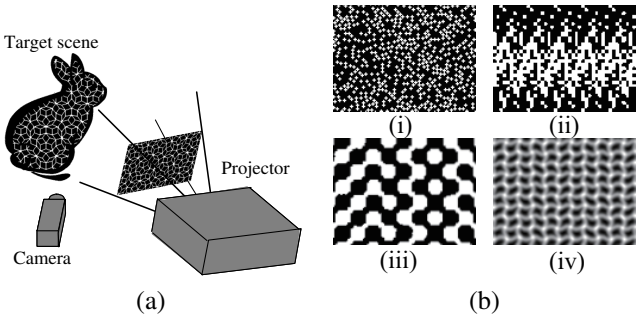


Fig. 1. (a) System configuration of projector and camera system. (b) Several patterns for one-shot scanning. (i) random dots, (ii) [22], (iii) [23] and (iv) [24].

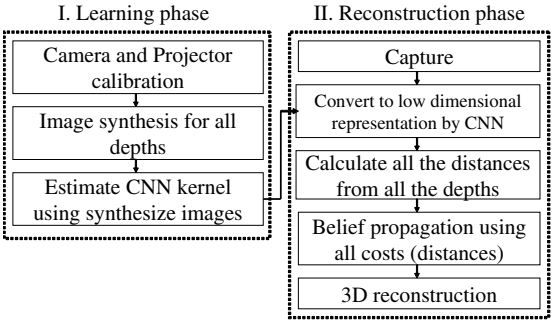


Fig. 2. Flow chart of our shape reconstruction algorithm.

B. Algorithm Overview

Our method has two phases, pattern learning and 3D reconstruction, as shown in Fig 2. In the pattern-learning phase, first, the projector and camera are calibrated. Then, through the use of calibration parameters, virtual images are synthesized by assuming that a planer board is placed at a specific depth, and the pattern is projected onto the board from a virtual projector and captured using a virtual camera. Those images are synthesized and stored by changing the depth, orientation, and material of the board. Finally, parameters and kernels of a two-layer CNN are estimated using a deep-learning framework to map the input image to low-dimensional values, typically ten-dimensional, in our case.

In the 3D-reconstruction phase, the input image is converted to a low-dimensional representation by applying a trained CNN with estimated parameters. Then, at each position of the captured image, distances from all of the depths are calculated and stored in a cost volume used to apply belief propagation (BP) in the next step. Once the depth values of all of the pixels are estimated, 3D shapes are recovered using the camera and projector calibration parameters.

C. Low dimensional representation for matching

In our method, images are mapped/encoded into a low dimensional representation. If the surface around a point $\boldsymbol{p}$ is assumed to be a local plane, then the homography H for

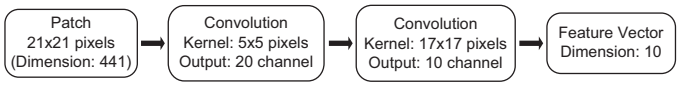


Fig. 3. The CNNs consist of two layers for encoding the local pattern. The input patch is 21×21 in this paper, and the dimension of the output feature vector is ten.

calculating the corresponding point of $\boldsymbol{p}$ is determined by the plane parameters, the orientation and depth of the plane. In this paper, the intensity of an input image at the point $\boldsymbol{p}$ is modeled as

$$I(\boldsymbol{p}) = aP(H\boldsymbol{p}) + b(\boldsymbol{p}) + c, \tag{1}$$

where $P(H\boldsymbol{p})$ is the intensity of the pattern image at the point $H\boldsymbol{p}$. The parameters a, b, and c represent the pattern brightness, the pixel noise of the imaging sensor, and the ambient brightness, respectively. Therefore, the input patch around the point $\boldsymbol{p}$ varies according to the parameters H, a, b, and c. Although other sources of variation can be considered, such as texture in a patch, focusing, and subsurface scattering, they will be considered in future work.

Encoding the patch to a low-dimensional representation must be stable with respect to the variation. One typical method for comparing local patches is the sum of squared distances (SSD). Comparison using SSD depends on the orientation of the local surface of the object, since the pattern reprojected to the camera image deforms according to the orientation. Therefore, it must calculate the SSD between the camera image and the projected pattern multiple times by changing the assumed orientation to improve the robustness of comparison. This increases the computational cost to find the correspondence between the camera image and the projected pattern. NCC or similar methods are known to perform better, but they typically require greater computational cost and are still not sufficient to compensate for variations.

The goal of mapping the projected pattern is to generate a low dimensional representation to reduce the computational cost to find the correspondence, which involves decoding the encoded pattern. In this paper, we propose a method for mapping/encoding the projected pattern to low-dimensional representation, which absorbs the above variation to compare image patches stably by low computational cost.

D. Learning the encoding model

The proposed method learns the encoding model specific to a given projected pattern. As described, we assume the projector-camera system is already calibrated. The training samples for learning are generated by reprojecting the pattern to the camera image by using the calibration parameters.

The encoding model proposed in this paper is based on CNNs. The network consists of layers, as shown in Fig.3. We assume that the pattern is locally encoded, and consider the distribution of intensity in a patch, whose size is 21×21 pixels in this paper. Since the projecting pattern typically consists of alphabet symbols encoded in a small area in the patches that are distinguishable from each other, the first layer is a convolutional layer with small window size, which is 5×5 pixels in this paper, to detect the structure of alphabet symbols. Since the symbols are detected by applying the convolution, it is not necessary to encode the pattern explicitly. An implicit encoding such as random dots can be used.

The codes of the pattern consist of combinations of symbols. Therefore, the second layer detects combinations of features calculated by the first layer. The output is the feature vector used for comparing patterns as the result of encoding. Because the number of patterns occurring in a patch is much smaller than the degrees of freedom of intensities in the patch, the dimension of the output can be much lower than that of the patch (10 in this paper). The rectified linear unit function is used as the activation function of the first layer.

Parameter estimation is based on the standard framework, mini-batch gradient descent, for estimating a neural network model with adaptive hyper-parameters [25]. The parameters are optimized by iterating the following steps.

- Randomly select multiple points in the camera image, and calculate the corresponding homographies by assuming the orientation and depth of surface planes.
- Reproject the pattern to the patches around the points, and add random pixel noises b to the patches in Eq.(1).
- Normalize the intensity for each patch using the mean and standard deviation of the patch. The pattern and ambient brightness, a and c in Eq.(1), are canceled by normalization.
- Construct a mini-batch of training samples and apply stochastic gradient descent to update the parameters.

In this paper, we design the feature vector so that the Euclidean distance between two features is equal to the SSD of the original patches that have no variation with respect to the projective distortion and pixel noise. Therefore, the loss function for estimating the parameters is defined as follows.

$$\min_f (\| f(\boldsymbol{x}) - f(\boldsymbol{y}) \|^2 - \| \boldsymbol{x}_0 - \boldsymbol{y}_0 \|^2)^2, \tag{2}$$

where f is the function of dimensional reduction by the neural network. $\boldsymbol{x}$ and $\boldsymbol{y}$ are two patches generated by adding variation, and $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ are the patches generated without variation.

E. Decoding input image using encoded features

The next step is decoding the input image to find the correspondence between the camera and projector. Since the pattern is encoded implicitly, the correspondence is provided by finding the best match of feature vectors along the epipolar line. Since the features of the projector pattern are independent of the camera image, they are calculated in advance of the decoding step. The matching cost for each pixel is calculated using the following steps.

1) Calculate the local mean and standard deviation for the patch, and normalize the intensity $\boldsymbol{x}$.
2) Apply the function of dimensional reduction to calculate the feature vector $f(\boldsymbol{x})$.
3) Assume a depth for the pixel and calculate the corresponding point in the projector pattern. The feature

vector of the corresponding point $f(\boldsymbol{y})$ is determined by bilinear interpolation of the pre-calculated feature of the projector pattern.

4) Calculate the cost $C = \| f(\boldsymbol{x}) - f(\boldsymbol{y}) \|^2$
5) Iterate the above steps 3 and 4 by changing the assumed depth.

Although the correspondence can be provided by selecting the minimum cost of the matches, some of the projector patterns do not provide a unique correspondence along the epipolar line. In such cases, a spatial constraint is added to determine the correspondence robustly similarly to the case of passive stereo methods. The proposed method applies the constraint based on a Markov random field (MRF) model. The matching cost for all pixels is defined as follows.

$$\sum_{\boldsymbol{p}} C(d_{\boldsymbol{p}}) + \lambda \sum_{\boldsymbol{p},\boldsymbol{q}} |d_{\boldsymbol{p}} - d_{\boldsymbol{q}}|, \tag{3}$$

where $C(d_{\boldsymbol{p}})$ is the cost if the depth at the point $\boldsymbol{p}$ is $d_{\boldsymbol{p}}$. The points $\boldsymbol{p}$ and $\boldsymbol{q}$ are neighboring pixels in the camera image. λ is a user-defined weight. The cost is minimized by iterative computation based on belief propagation [26]. The depth of the minimum cost after minimization is selected as the best match.

F. Detecting pixels without patterns

In some cases, there are pixels in a camera image that the pattern is not projected onto. Those pixels must be omitted from the depth estimation because they have no information for finding correspondence. Although a simple method for finding them is thresholding by brightness, this is not applicable if the pattern brightness is low or if the ambient brightness is high. The proposed method uses the following criteria to discriminate pixels with and without patterns. If the following condition is satisfied, the pixel is regarded as one that the pattern is projected onto it.

$$\max_{d_{\boldsymbol{p}}} C(d_{\boldsymbol{p}}) / \min_{d_{\boldsymbol{p}}} C(d_{\boldsymbol{p}}) > \theta, \tag{4}$$

where $C(d_{\boldsymbol{p}})$ is the cost after MRF minimization and θ is a user-defined threshold. After thresholding, morphological operations are applied to remove small regions.

IV. Experiments

In the experiments, we set up a projector-camera system using an off-the-shelf video projector, as shown in Fig.4. The resolutions of the camera and projector are 1600×1200 pixels and 1024×768 pixels, respectively. Since the proposed method has no limitations on the projector pattern, it can theoretically be applied to monochrome, color, and time-multiplexing patterns, but we focus on monochrome fixed patterns as shown in Fig. 1(b). Therefore, only a single channel of the input data is used to reconstruct the 3D shape of a target object.

The proposed method uses the imaging model of Eq.(1) to estimate the function of dimensional reduction. Since it is expected that the proposed method is robust with respect to the noise of the input image, the noise model is introduced in the training data set. In the first experiment, we test the robustness of the proposed system in the case of a low signal-noise (S/N) ratio.

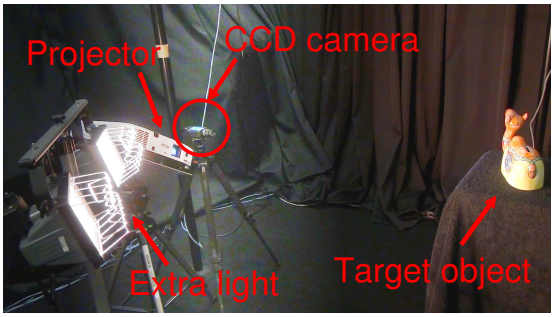


Fig. 4. Experimental system with a camera and a video projector. External lights are used as ambient light source.

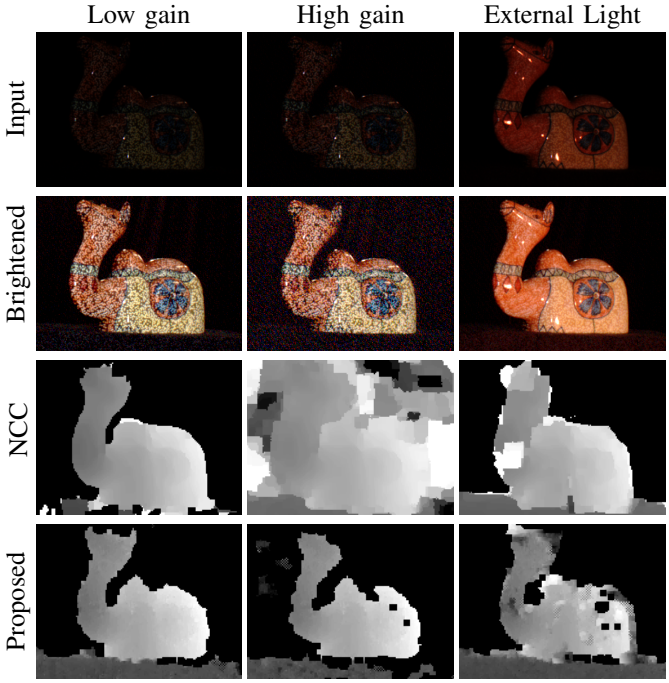


Fig. 5. The proposed method is compared to an NCC-based method in the case of low S/N images. The images in the top row are input images. The images in the second row are brightened from the input images. The third and bottom row show the depthmaps using the NCC-based method and the proposed method, respectively. The object is placed on a black cloth, which is reconstructed successfully using the proposed method.

Under a wide variation of brightness change of an input image, NCC is often used to compare image patches, and thus, we compare the proposed method to a method that calculates the matching cost using NCC. MRF-based energy minimization is applied to find the correspondence using the NCC data cost and the regularization cost between adjacent pixels. Fig.5 shows the results for the images of a low S/N ratio. A random dot pattern is projected to the object in this experiment. The leftmost column shows the input images. The images of the left column have low gain and are dark. In the

TABLE I

THE RMS ERRORS ARE CALCULATED BY COMPARING THE RESULTS OBTAINED USING THE FOUR METHODS AND THE GROUND TRUTH, WHICH IS OBTAINED BY THE TIME-MULTIPLEXING 3D SCANNING METHOD THAT USES GRAY CODES. THE UNIT IS MILLIMETERS.

Methods	Boots	Hand	Skull	Lizard	Average
Wave-oneshot	3.61	1.76	5.83	13.28	6.12
NCC/dot	5.91	2.78	7.64	3.66	5.00
NCC/wave	6.70	3.85	7.00	9.18	6.68
PCA/dot	11.92	9.91	13.01	11.66	11.63
PCA/wave	6.59	8.12	8.84	13.41	9.24
Proposed/dot	3.32	2.05	3.44	4.13	3.24
Proposed/wave	2.76	5.33	6.07	3.04	4.30

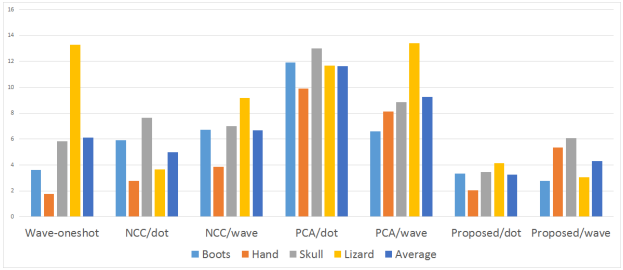
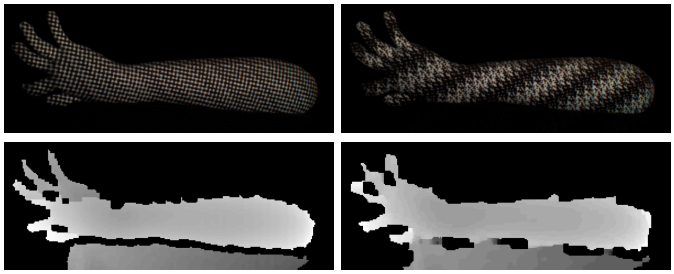


Fig. 7. The RMS errors are compared for four methods. The proposed method shows robust results for all cases compared to the other methods.



(a) (b)

Fig. 8. Calculating depthmaps by projecting two patterns: (a) PN-sequence grid pattern [23], (b) density modulated binary pattern [22]

center column, the image is captured with high gain. In the right column, the object is illuminated using external lights as an ambient light source. The images in the second row are brightened from the input images by color correction. The object is placed on a black cloth and the pattern is also projected onto the cloth. The figures in the third row are the depthmaps obtained using the NCC-based method. The darker pixels indicate closeness to the camera. The low intensity pixels after normalization are regarded as background and discarded from reconstruction. The reconstruction works well in the case of low gain, but the background detection fails in the case of high gain. The reconstruction of the image with external light fails around the neck. In the bottom row, the proposed method succeeds in reconstructing the shapes of both the object and the black cloth even if the S/N ratio is low. While the reconstruction is degraded in the case of high gain and external light, the background is correctly detected and depth is obtained for most of the object.

Next, we evaluate the accuracy of the proposed method by projecting two different patterns, random dots and wave grid [24], for one-shot 3D scanning in this experiment. The top row in Fig.6 shows the input images of four objects. The random dot and wave grid patterns are projected on the left and right sides, respectively. The middle and bottom rows show the depthmaps and meshes, respectively using the proposed method. The proposed method is compared to three other methods, Wave-oneshot [24], a NCC-based, and a PCA-based method. The Wave-oneshot method uses a wave-grid pattern to reconstruct a 3D shape. The NCC-based method is the same as the one in the above experiment. The PCA-based method is a learning-based approach that extracts features using PCA. It calculates the eigenvectors of the patches synthesized from the projector pattern, and chooses ten vectors with large eigenvalues. These eigenvectors are convolved with the input image as the kernel, and the ten convolutions are used as the feature vector for matching. In particular, the PCA-based method is a linear method of extracting features for comparison. MRF energy minimization is also applied to the PCA-based method after calculating the data cost as the Euclidean distance using the feature vectors.

To evaluate the accuracy, we compared the 3D shapes obtained by four methods, Wave-oneshot [24], NCC-based, PCA-based, and the proposed method. The input images are captured by changing the conditions, as shown in Fig.5. The errors of depthmaps are calculated using the root-mean-square (RMS) errors from the ground truth, which is obtained by the time-multiplexing 3D scanning method that uses Gray codes. The background including the black cloth is discarded from the comparison. Table I and Fig.7 show the RMS error results. The unit of the values is millimeters. The errors of the PCA-based method are larger than those of other methods, which are considered to be the result of the linear approach being not sufficient to absorb the variation of input images. The Wave-oneshot method basically shows good results, but the error is large in the case of Lizard, which has a high frequency texture and detecting lines is difficult. The results of the NCC-based method are generally better than those of the PCA-based method, and this shows best in the random dot patterns for Hand and Lizard. However, the proposed method results in low RMS errors for all cases and is, consequently, the most robust of these methods.

Next, we tested 3D reconstruction using the proposed method by projecting other patterns. Fig.8 shows the input images and the depthmaps that result from projecting two different patterns. Pattern (a) is a chess-board grid pattern encoded based on a pseudo-noise (PN) sequence [23]. Pattern (b) is density-modulated binary patterns [22], combining window matching and phase-shift approach. Even though the proposed method does not have explicit knowledge of the patterns, it succeeded in calculating the depthmaps by using these patterns, proving that the proposed method provides a universal approach for various patterns of the active stereo methods.

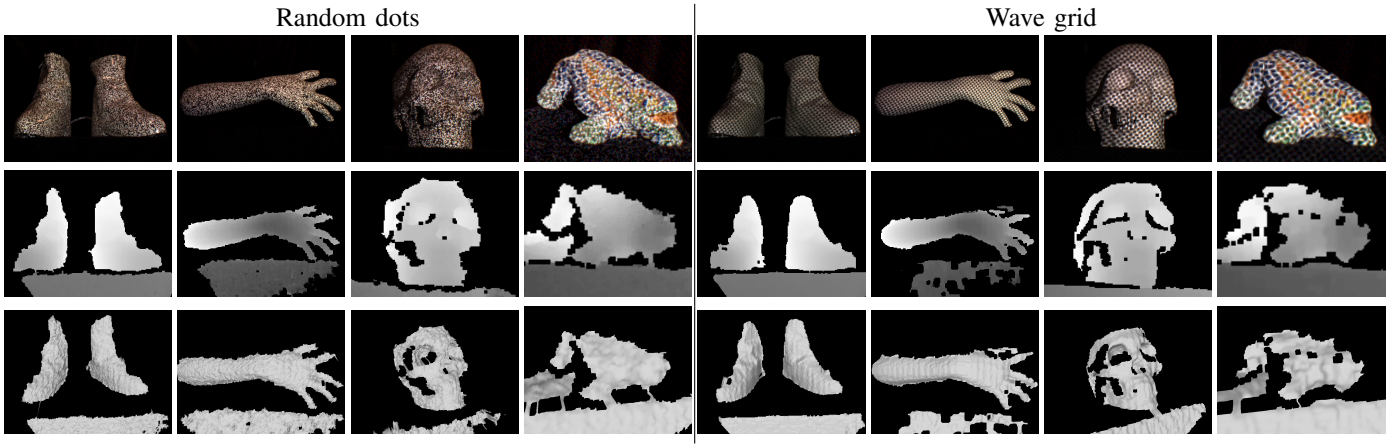


Fig. 6. Two different patterns, random dots and wave grid, are projected onto the objects. The depthmaps and meshes are generated by the proposed method for four objects, Boots, Hand, Skull, Lizard. These objects are placed on a black cloth.

V. Conclusion

In this paper, we proposed a universal decoding algorithm based on a learning approach. Parameters of a mapping function from a raw image to a low-dimensional representation were efficiently trained using a deep learning technique. With our technique, since critical features of the pattern are automatically extracted from a huge database that is synthesized by using a virtual projector and camera, no manual intervention or customized algorithm is necessary, greatly helping to realize learning based approach without actually preparing huge image data set. Experimental results were shown to demonstrate that our technique is stable, irrespective of the target object material, lighting conditions, or sensor noise, compared to existing methods. In the future, real-time implementation on GPU will be important.

References

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

[2] ——, "Kinect for Windows," 2013, http://www.microsoft.com/en-us/kinectforwindows.

[3] Mesa Imaging AG., "SwissRanger SR-4000," 2011, http://www.swissranger .ch/index.php.

[4] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern recognition*, vol. 43, no. 8, pp. 2666–2680, 2010.

[5] R. Benveniste and C. Ünsalan, "A color invariant based binary coded structured light range scanner for shiny objects," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 798–801.

[6] Q. Li, M. Biswas, M. R. Pickering, and M. R. Frater, "Accurate depth estimation using structured light and passive stereo disparity estimation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 969–972.

[7] X. Zhang, Y. Li, and L. Zhu, "Color code identification in coded structured light," *Applied optics*, vol. 51, no. 22, pp. 5340–5356, 2012.

[8] S. Tang, X. Zhang, and D. Tu, "Fuzzy decoding in color-coded structured light," *Optical Engineering*, vol. 53, no. 10, pp. 104 104–104 104, 2014.

[9] Q. Li, M. Biswas, M. R. Pickering, and M. R. Frater, "Dense depth estimation using adaptive structured light and cooperative algorithm," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011, pp. 21–28.

[10] H. Kawasaki, R. Furukawa, , R. Sagawa, and Y. Yagi, "Dynamic scene shape reconstruction using a single structured light pattern," in *CVPR*, June 23-28 2008, pp. 1–8.

[11] A. O. Ulusoy, F. Calakli, and G. Taubin, "Robust one-shot 3d scanning using loopy belief propagation," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 15–22.

[12] V. Couture, N. Martin, and S. Roy, "Unstructured light scanning to overcome interreflections," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1895–1902.

[13] H. Kawasaki, H. Masuyama, R. Sagawa, and R. Furukawa, "Single colour one-shot scan using modified penrose tiling pattern," *IET Computer Vision*, vol. 7, no. 5, pp. 293–301, October 2013.

[14] H. Kawasaki, S. Ono, Y. Horita, Y. Shiba, R. Furukawa, and S. Hiura, "Active one-shot scan for wide depth range using a light field projector based on coded aperture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3568–3576.

[15] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[16] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition," *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[17] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International journal of computer vision*, vol. 14, no. 1, pp. 5–24, 1995.

[18] M. Uenohara and T. Kanade, "Use of fourier and karhunen-loeve decomposition for fast pattern matching with a large set of templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 8, pp. 891–898, 1997.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1592–1599.

[21] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[22] Z. Yang, Z. Xiong, Y. Zhang, J. Wang, and F. Wu, "Depth acquisition from density modulated binary patterns," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 25–32.

[23] P. Vuylsteke and A. Oosterlinck, "Range image acquisition with a single binary-encoded light pattern," *IEEE Trans. on PAMI*, vol. 12, no. 2, pp. 148–164, 1990.

[24] R. Sagawa, K. Sakashita, N. Kasuya, H. Kawasaki, R. Furukawa, and Y. Yagi, "Grid-based active stereo with single-colored wave pattern for dense one-shot 3D scan," in *Proc. 2012 Second Joint 3DIM/3DPVT Conference*, Zurich, Switzerland, Oct. 2012, pp. 363–370.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[26] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *IJCV*, vol. 70, pp. 41–54, 2006.