Denser Feature Correspondences for 3D Reconstruction

Fairuz Safwan Mahad[1,a)] Masakazu Iwamura[1,b)] Yuzuko Utsumi[1,c)] Koichi Kise[1,d)]

Abstract: The 3D reconstruction of an object is heavily dependent on the amount of texture on the surface of the object and the correspondences acquired between images from different viewpoints. Objects with lack of texture affects the reconstructed 3D model degrading its quality and completeness. This occurs due to the SIFTs and Harris's incapability of extracting sufficient features from regions with less texture thus resulting in a lack of correspondences in such regions. Gaps occurs in the reconstructed 3D model due to the lack of correspondences. The paper aims to obtain more correspondences through the implementation of convolution in conjunction with a set of randomized kernels. The use of randomized kernels changes the structure of the images in different ways allowing SIFT to extract features in less textured regions which could not be realized conventionally. Experimental results showed that the proposed method is able to fill up more gaps improving the completeness of the 3D model of less textured objects.

1. Introduction

There is an increasing demand in three-dimensional (3D) models especially in industries revolving the virtual environment. The use of 3D models has become a necessity in the rapid growth of technologies. There are several ways to extract the 3D information out of an object. Furthermore, it is particularly divided into two distinct categories, namely the active methods (c.f., [10], [1], [13], [8], [5], [6], [16]) and the passive methods (c.f., [11], [3], [2], [15]).

The active methods describe the exploitation of light sources under a specially controlled illumination. It usually realized in conjunction with other additional equipments such as cameras and projectors in order to reconstruct a 3D model of an object. While the active method show promising results, the main drawback is that the light sources as well as the illumination have to be specially controlled resulting in a flexibility constraint. The passive methods on the other hand offer much more flexibility. It does not concern with any specially controlled illumination instead it only concerns with the input images. As agreed by [9] in comparing active and passive methods, active methods tend to be less demanding due to the enviromental settings constraint which do not offer flexibility. In addition, [9] evaluates the 3D reconstruction from multiple and uncalibrated images such as the structure-from-motion technique to be one of the most promising 3D reconstruction techniques. Therefore, this technical report will focus on passive 3D reconstruction methods here onwards.

Furukuwa et al. [3] proposed a 3D reconstruction method called dense multi-view stereo which assumes cameras are calibrated (i.e., camera parameters are known). It is categorized into the passive methods which involve a collection of images from different viewpoints. [3] produces promising results for objects with rich textures but not for objects with less texture. The method uses Scale Invariant Feature Transform (SIFT) [7] and Harris corner detector [4] (hereafter, referred as Harris) to detect features. The completeness of the 3D reconstructed model relies heavily on the correspondences between features from different viewpoints. A lack of correspondences results in an incomplete 3D reconstructed model where gaps will occur. SIFT and Harris could not extract features from regions with less textures which therefore caused the occurrence of the gaps.

Due to the above-mentioned drawback of SIFT, [12] proposed a method which deals with texture-less objects. Their method is called Bunch of Lines Descriptors (BOLD), exploiting edges in order to detect texture-less objects. The BOLD method extracts edges and segments mainly from contours of objects and utilizes them as descriptors. Despite being robust to occlusion, clutter and scalability in recognizing texture-less objects, the BOLD method is capable of recognizing artificial objects such as bowls, cups and tools but not natural objects such as plants

This technical report proposes a method to extract features from regions with less texture. The aim of the research is to acquire as many features and correspondences as possible from textured regions as well as regions with less texture for preventing gaps from occurring. In order to acquire features from less textured regions, the proposed method introduces the use of convolution. Convolution is capable of increasing the complexity of the images. In order for convolution to work, a kernel has to be used. Designing a specific kernel for the sole purpose of extracting features from less textured regions is not an easy task. Therefore the proposed method implements a set of randomized kernel. Implementing convolution in conjunction with a set of randomly generated kernel randomly alters the structure and complexity of images in many different ways. This helps in obtaining more SIFT and Harris features in less textured regions. Figure 1 il-

[1] Graduate School of Engineering, Osaka Prefecture University 1-1, Gakuencho, Naka, Sakai, Osaka 599-8531, Japan
[a)] safwan@m.cs.osakafu-u.ac.jp
[b)] masa@cs.osakafu-u.ac.jp
[c)] yuzuko@cs.osakafu-u.ac.jp
[d)] kise@cs.osakafu-u.ac.jp

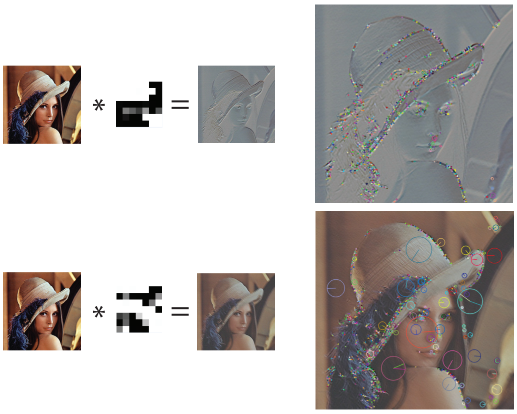


Fig. 1: A comparison of extracted SIFT features between 2 randomized kernels.

lustrates two examples of convoluted images and also its features extracted by SIFT. The original images are convoluted with different randomized kernels. By observing the extracted features of both convoluted images, it can be seen that the detected features of both convoluted images mostly differ. This demonstrates the capability of convolution in conjunction with a set of randomized kernels in an effort to extract as many features in many different parts of the image as possible.

2. Related Work

2.1 Active Methods

There are numerous researches which aims to improve 3D reconstruction using several different approaches either by using the active methods or the passive methods. The active methods involves the use of light sources such as lasers in conjunction with other additional equipments in order to extract the 3D information of the intended object. [10] proposed an active 3D reconstruction method which uses a camera and a projector. Assuming the camera and projector are both calibrated, a grid pattern of horizontal and vertical lines is projected from the projector onto the intended object and the lines are captured by the camera. These captured lines are then analyzed to have its 3D information extracted. On top of that, [1] and [13] proposed active 3D reconstruction methods utilizing laser rays as their light source. On the other hand, [8], [5], [6] and [16] proposed a photometric stereo approaches which are also active based 3D reconstruction method. The photometric stereo method are able to reconstruct a 3D model of an object despite the lack of texture. However, the photometric stereo method only can only be realized in dark areas where light is not present and is not possible to be realized outdoors or anywhere where light is present. Despite the promising results of the active methods, it requires environment to be specially controlled unlike the passive methods which offer much more flexibility.

2.2 Passive Methods

Snavely et al. [11], for instance, proposed a state-of-the-art Structure-From-Motion (SfM) based system called Bundler. Similarly to the method proposed by Furukawa et al. [3], the method proposed by Snavely et al. [11] is a passive based method which involves a collection of images from different viewpoints. The method uses SIFT [7] to detect features in order to find correspondences between images. However, the completeness of the reconstructed 3D representation relies heavily on having correspondences between images. With few correspondences, there will be gaps in the reconstructed 3D representation. The SfM method estimates the 3D points from the two dimensional (2D) points in images from different viewpoints by using the correspondences of features in the images. Figure **??** and **??** depicts an example of where correspondences are obtained and otherwise. As illustrated in Figure **??**, a 2D point on an image has to correspond to a 2D point in another image from another viewpoint. Figure **??** on the other hand illustrates a case where a 2D point in an image does not correspond to any 2D point in neither images. As a result of this situation, gaps will occur in the reconstructed 3D model degrading its quality. This demonstrates how vital the corresponding points are in 3D reconstruction. The method also requires location information and intrinsic parameters such as the focal length in order to reconstruct a 3D representation. Such information can be extracted from the EXIF tag in images. SfM does not require a calibrated camera meaning it does not require camera parameters instead it is able to estimate the camera parameters as one of its output. This feature offers flexibility where constraints are set loose and its 3D reconstruction approach is much easier to implement.

Another passive based 3D reconstruction method is the multi-view stereo. Furukawa et al. [3] proposed a multi-view stereo method. Unlike the method of Snavely et al. [11], Furukawa et al. [3] assumes cameras are calibrated (i.e., camera parameters are known). Therefore in order for it to work, it requires camera parameters as inputs. The multi-view stereo aims to reconstruct a 3D representation of an object. The state-of-the-art multi-view stereo method proposed by [3] called PMVS2 is a novel calibrated multi-view stereo algorithm. The method of [3] is a dense version of multi-view stereo which produces dense points. Similarly to Snavely et al. [11], Furukawa et al. [3] requires corresponding points in order to reconstruct the 3D representation of the object or scene. These highlight the importance of having corresponding points in 3D reconstruction. In additional, Wu et al. [15] also proposed a dense multi-view stereo based 3D reconstruction method which is actually an improvement based on the method of [3]. Using the same pipeline as [3], [15] takes in a collection of images and improves the propagation process in order to yield a much more complete 3D reconstructed model as compared to [3] under certain circumstances. Despite the improvements, the computational cost is immensely expensive. Another passive 3D reconstruction method proposed by [2] differs slightly in terms of its approach. [2] proposed a passive 3D reconstruction method which uses silhouette and texture information extracted from images. These information is then fused in order to reconstruct a 3D model.

The proposed method adopts the conventional pipeline which is using [11]'s method in conjunction with the method of [3]. Figure 2 depicts the summary of processes and components involved in the conventional method's pipeline. A typical pipeline is to use SIFT to detect features and acquire correspondences and use

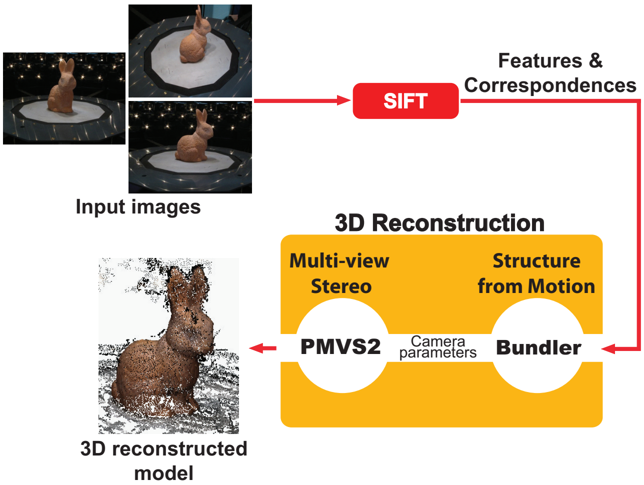


Fig. 2: Summary of the Conventional Method.

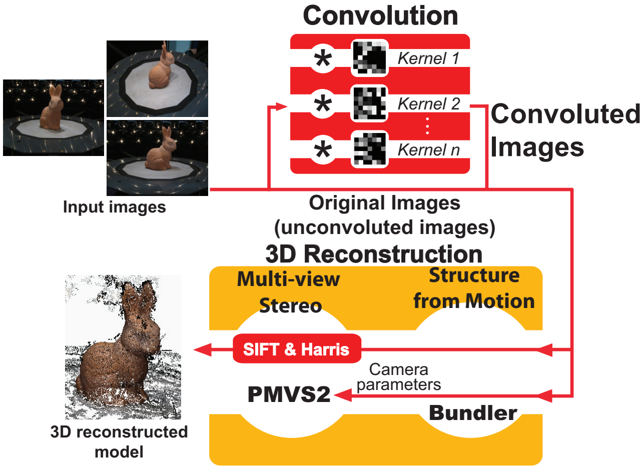


Fig. 3: Overview of the Proposed Method.

the acquired features and correspondences as input to Bundler. Bundler produces estimated camera parameters which will be subsequently used as input to PMVS2. PMVS2 will then be used to reconstruct a 3D structure of the object. In the conventional method, the reconstructed 3D structure suffers from many visible gaps due to the lack of correspondences acquired especially in regions with lack of texture. Therefore, the aim of this research is to acquire more correspondences in order to fill up as many gaps as possible and acquire a much more complete 3D reconstructed model.

3. Proposed Method

The proposed method adopts the conventional method's pipeline and on top of that, introduces the implementation of convolution. The two key factors to the proposed method are the implementation of convolution and also the use of a set of randomized kernels. In order to obtain features from less textured regions, the proposed method implements convolution. Input images are convoluted with a randomly generated kernel for each run. The randomize kernels are used to increase the complexity of the images. This allows more features to be obtained from less textured regions which could not be realized in the conventional method.

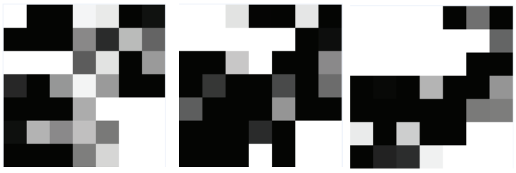
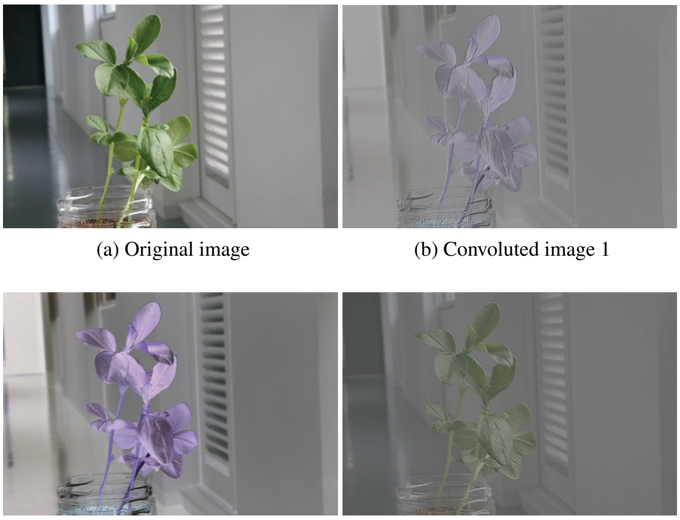


Fig. 4: Examples of Randomized Kernels.



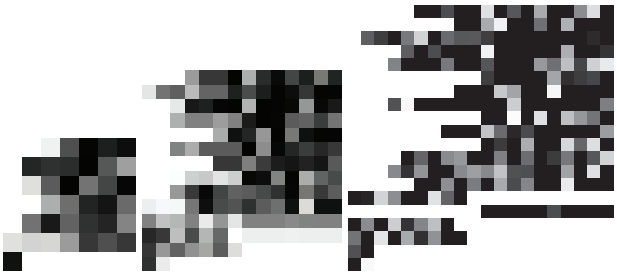
(a) Original image
(b) Convoluted image 1
(c) Convoluted image 1
(d) Convoluted image 1

Fig. 5: Examples of original (a) and convoluted images (b)-(d).

Figure 3 illustrates an overview of the proposed method. Bundler is used to acquire estimated camera parameters based on original input images. These estimated cameras are used in PMVS2 as an input. In PMVS2, SIFT and Harris features are detected from the original image set and also the convoluted image sets. The features detected are merged before performing the initial feature matching in PMVS2. For the case of using 10 randomized kernels, there will be a total of 11 sets of images involved. This includes 10 convoluted image sets where each image set is convoluted with a different randomly generated kernel and also the original set of input images which are not convoluted. Assuming there are 15 images in each set. Therefore, it involves a total of 165 images of which 150 are convoluted images and the remaining are original images which are not convoluted.

3.1 Randomized Kernel

Convolution requires a kernel in order to work. Some researchers have pre-determined a specific kernel which best suits their research purpose. For instance, [14] designed a kernel for the purpose of de-blurring natural images. However, the mentioned kernel applies only for de-blurring natural images and not for extracting features. Therefore, the proposed method adopt the use of randomly generated kernels. Figure 4 illustrates few examples of the randomized kernels produced by the proposed method. All elements in the kernel's matrix are randomized with floating numbers in the range between 0 to 10. These values are then normalized so that the sum of elements in a kernel is in the range of -1 and 1. Figure 5 shows the comparison of images convoluted with different randomized kernels. Figure 5a shows an



(a) 7x7 pixels (b) 14x14 pixels (c) 20x20 pixels

Fig. 6: Examples of kernel sizes.

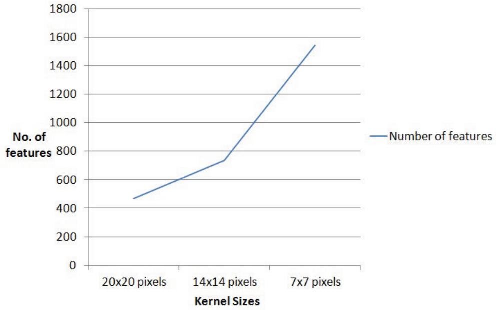


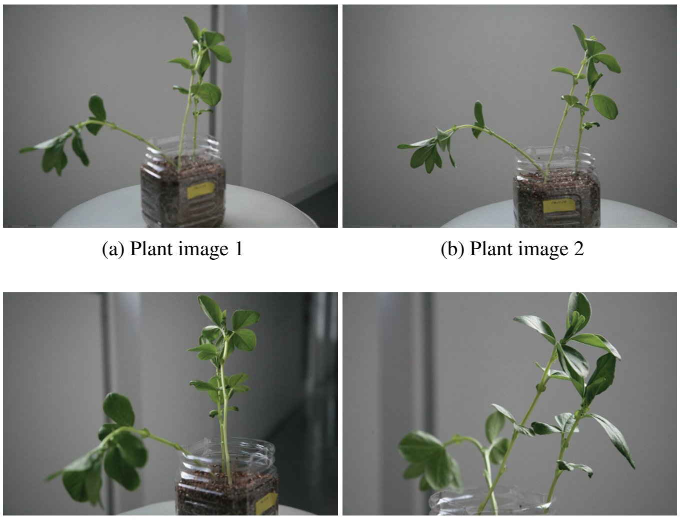
Fig. 7: Relationship between No. of features extracted and kernel size.

example of the original image before the implementation of convolution while Figures 5b, 5c and 5d show the convoluted images produced by convoluting the original image with the randomized kernels in Figure 4 respectively.

Figure 6 illustrates several sizes of the kernel of which are 7×7 pixels in Figure 6a, 14 × 14 pixels in Figure 6b and 20 × 20 pixels in Figure 6c. The size of the kernel directly affects the result of the proposed method. There is a relationship between the size of the kernel and the number of extracted features. It is discovered that the smaller the size of the kernel, the higher the probability of extracting more features. Therefore, the proposed method uses the kernel of size 7x7 pixels. Experimental data of various kernel sizes are presented in the following chapter.

4. Experiments

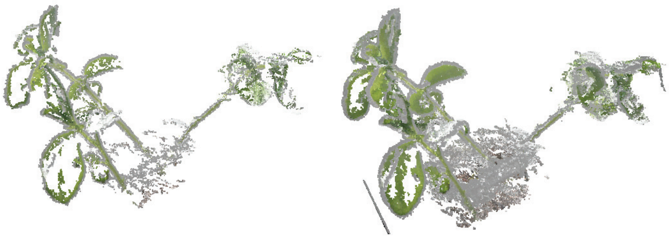
As mentioned, there is a relationship between the size of the kernel and the number of extracted features. In most cases the smaller the size of the kernel the higher the number of features extracted. Figure 7 shows the numerical data comparing the number of features extracted from images convoluted by three different sizes of kernels which are 20x20 pixels, 14x14 pixels and 7x7 pixels. It shows that the kernel of sized 7x7 pixels has more features extracted as compared to the other two sizes. The inclining graph also further shows the relationship between the size of the kernel and the number of features extracted. Although it does not necessarily apply on all cases, a smaller sized kernel increases the probability of acquiring more features and also correspondences. Therefore, the kernel of size 7x7 pixels was used in the proposed method and also in the experiments presented in the following section.



(a) Plant image 1 (b) Plant image 2

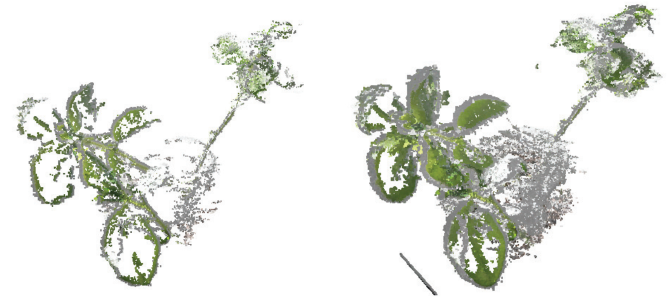
(c) Plant image 3 (d) Plant image 4

Fig. 8: Examples of plant dataset.



(a) Conventional Method (b) Proposed Method

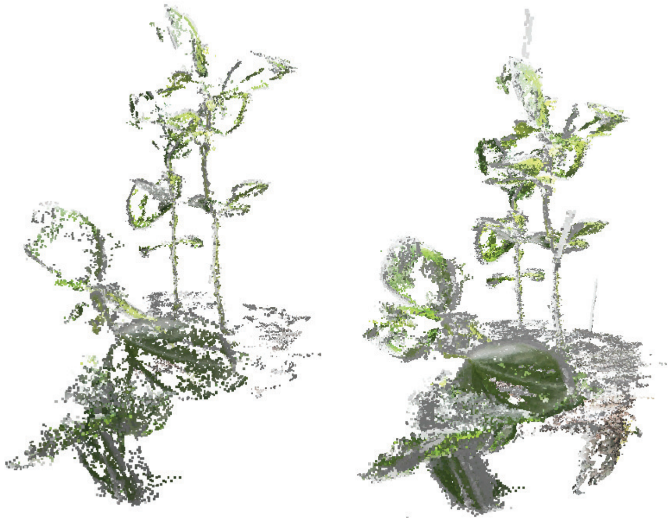
Fig. 9: Experiment (View 1) with (a) conventional and (b) proposed method.



(a) Conventional Method (b) Proposed Method

Fig. 10: Experiment (View 2) with (a) conventional and (b) proposed method.

An experiment has been conducted with a plant dataset of 27 images. Figure 8 shows several examples of the plant dataset from different viewpoints. The number of randomized kernels used for the experiment was 100. The dataset was convoluted 100 times with 100 different randomly generated kernels. This involved a total of 2727 images consisting of 2700 convoluted images and 27 original images. After combining the correspondences of all 101 sets of images including the original dataset which was not convoluted, the 3D representation was reconstructed using the method of [11] followed by the method of [3]. Figures 9 to 11 shows the results of (a) being the conventional method and (b) being the proposed method. The results shows



(a) Conventional Method (b) Proposed Method

Fig. 11: Experiment (View 3) with (a) conventional and (b) proposed method.

that the (b) proposed method was able to fill up more gaps as compared to the (a) conventional method. More patches can be seen filling up the areas in the leaves and especially some main veins of the leaves are reconstructed. There are also some outlines of the leaves which are able to be reconstructed in the (b) proposed method but was not able to be reconstructed in the (a) conventional method. These also means that there are more correspondences acquired in the (b) proposed method in comparison with the (a) conventional method.

5. Concluding Remarks

The 3D reconstruction of an object or scene relies heavily on the acquired correspondences between images in different viewpoints. In the conventional method, there will be gaps in the reconstructed 3D representation if there are no correspondences at a particular point. The aim of the research is to obtain more correspondences in order to improve the completeness of the 3D reconstructed representation of the object or scene. SIFT is not able to extract features from less textured regions which serves as one of the reason for the lack of correspondences. Therefore, the proposed method uses convolution in conjunction with a set of randomized kernels in order to increase the complexity of the image whilst increasing the probability of acquiring more correspondences. Experimental results clearly showed that the proposed method is able to cover up more gaps and fill more areas in the leaves as compared to the conventional method. Although the completeness is still lacking, the proposed method still shows progress and develops a much more complete 3D representation of the object.

References

[1] Espinal, J., Ornelas, M., Puga, H. J., Carpio, J. M. and Munoz, J. A.: 3D Object Reconstruction Using Structured Light and Neural Networks, *Proc. of Electronics, Robotics and Automotive Mechanics Conference (CERMA)*, pp. 74–79 (2010).

[2] Esteban, C. H. and Schmitt, F.: Silhouette and Stereo Fusion for 3D Object Modeling, *Comput. Vis. Image Underst.*, Vol. 96, No. 3, pp. 367–392 (2004).

[3] Furukawa, Y. and Ponce, J.: Accurate, Dense, and Robust Multiview Stereopsis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 8, pp. 1362–1376 (2010).

[4] Harris, C. and Stephens, M.: A combined corner and edge detector, *Proc. of 4th Alvey Vision Conference*, pp. 147–151 (1988).

[5] Hernandez, C., Vogiatzis, G. and Cipolla, R.: Multiview photometric stereo, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 3, pp. 548–554 (2008).

[6] Lim, J., Ho, J., Yang, M.-H. and Kriegman, D.: Passive photometric stereo from motion, *Tenth IEEE International Conference on Computer Vision*, Vol. 2, pp. 1635–1642 (2005).

[7] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *IJCV*, Vol. 60, No. 2, pp. 91–110 (2004).

[8] Lu, F., Matsushita, Y., Sato, I., Okabe, T. and Sato, Y.: From intensity profile to surface normal: photometric stereo for unknown light sources and isotropic reflectances, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 10, pp. 1999–2012 (2015).

[9] Moons, T., Van Gool, L. and Vergauwen, M.: 3D Reconstruction from Multiple Images Part 1: Principles, *Found. Trends. Comput. Graph. Vis.*, Vol. 4, No. 4, pp. 287–404 (2010).

[10] Sagawa, R., Ota, Y., Yagi, Y., Furukawa, R. and Asada, N.: Dense 3D reconstruction method using a single pattern for fast moving object, *Proc. of 14th International Conference on Computer Vision (ICCV 2013)*, pp. 1779–1786 (2013).

[11] Snavely, N., Seitz, S. M. and Szeliski, R.: Photo Tourism: Exploring Photo Collections in 3D, *Proc. of ACM SIGGRAPH 2006*, pp. 835–846 (2006).

[12] Tombari, F., Franchi, A. and Stefano, L. D.: BOLD Features to Detect Texture-less Objects, *Proc. of 14th International Conference on Computer Vision (ICCV 2013)*, pp. 1265–1272 (2013).

[13] Usamentiaga, R., Molleda, J. and Garcia, D.: Structured-light sensor using two laser stripes for 3D reconstruction without vibrations, *Sensors*, Vol. 14, No. 11, pp. 20041–20063 (2014).

[14] Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A. and Tumblin, J.: Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocusing, *ACM Trans. Graph.*, Vol. 26, No. 3 (2007).

[15] Wu, T. P., Yeung, S. K., Jia, J. and Tang, C. K.: Quasi-dense 3D reconstruction using tensor-based multiview stereo, *Proc. of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1482–1489 (2010).

[16] Yuille, A. and Snow, D.: Shape and albedo from multiple images using integrability, *Proc. of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 158–164 (1997).