Depth from Defocus via Active Multispectral Quasi-random Point Projections using Deep Learning

Avery Ma, University of Waterloo, ON, Canada
Alexander Wong, University of Waterloo, ON, Canada
David A. Clausi, University of Waterloo, ON, Canada

Abstract

A novel approach for inferring depth measurements via multispectral active depth from defocus and deep learning has been designed, implemented, and successfully tested. The scene is actively illuminated with a multispectral quasi-random point pattern, and a conventional RGB camera is used to acquire images of the projected pattern. The projection points in the captured image of the projected pattern are analyzed using an ensemble of deep neural networks to estimate the depth at each projection point. A final depth map is then reconstructed algorithmically based on the point depth estimates. Experiments using different test scenes with different structural characteristics show that the proposed approach can produced improved depth maps compared to prior deep learning approaches using monospectral projection patterns.

1 Introduction

Depth measurement plays an important role in the understanding of a scene and depth-sensing cameras allow for recovering this vital information in many applications.

The use of active depth-sensing techniques has been gaining popularity due to its superior performance, efficiency, and ease of application, and a number of such techniques exist. For example, laser scanners are widely used in robotics and autonomous driving to recover accurate depth measurements [1]. Although such technologies have improved over the years, they remain very expensive and thus not feasible in a wide range of applications where there are stricter cost and complexity constraints. Stereo-based structured light systems are another popular approach due to lower cost and complexity, but require a fundamental trade-off between baseline and depth accuracy that makes it ineffective to utilize for in certain scenarios. Furthermore, while newer active depth-sensing technologies are constantly being advanced and address many of of the drawbacks of the previous methods, they nevertheless rely on specialized hardware that increase cost and complexity. As such, alternative active depth-sensing techniques that address these challenges are required.

In this paper, we present a novel approach for inferring depth measurements using a single camera by leveraging the idea of depth from defocus using multispectral active quasi-random point projections and deep learning. The proposed approach eliminates the need for a baseline (the projector can be completely in-line with the camera), has a relatively simple setup, and provides improved spatial resolution beyond what can be done with single-wavelength approaches [2, 3, 4], and would be expected to lead to very compact and low-cost active depth-sensing systems.

The novel method involves actively illuminating a multispectral quasi-random point pattern onto the scene of interest. The projected pattern is captured using a RGB camera, and an ensemble of deep neural networks is used to estimate point-wise depth based on the captured projected pattern. A computational reconstruction method is used to generated a final depth map from the sparse depth estimation results from the deep neural networks.

The paper is organized as follows. In Section 2, related work in the area of depth from defocus is discussed. In Section 3, the proposed depth inference method is described in detail, while in Section 4 the experimental results are presented. Finally, conclusions are drawn in Section 5.

2 Related Work

The concept of depth from defocus (DfD) has been explored in past literature, and a number of approaches have been proposed. Traditional DfD methods estimate depth by studying the difference in blurriness between two images that are captured at different focal lengths. Different filters have been previously proposed for determining the degree of blur [5, 6, 7]. A major limitation to such DfD approaches is that blur detection can be unreliable in a number of different situations, especially in untextured areas of the image. This problem faced by traditional DfD methods is mitigated in active structured light-based depth-sensing approaches, where an optical projection is used to find correspondences for triangulation. A review of structured light approaches for depth measurement is provided by Salvi *et al.* [8]. The key benefit of such structured light-based approaches is that they do not depend on the objects in the scene, and as such, they are particularly effective in untextured regions. In contrast, active structured light-based depth-sensing systems require a fundamental trade-off between baseline and depth accuracy that makes the method ineffective in certain scenarios. As such, we are motivated to leverage the strengths from both DfD and active depth-sensing methods to design a method that mitigates their individual limitations, thus enabling systems with a simple setup yet achieve reliable results in the depth measurements.

The concept of using DfD using active projections has also been explored in the literature. Pentland *et al.* [9] proposed the use of evenly-spaced line projections to determine depth based on line spread, and this simple method is able to achieve low-resolution depth maps. Nayar *et al.* [10] proposed the use of a dual sensor plane with an optimized projection and camera setup to produce a dense depth map while reducing front/back focal ambiguity. Ghita *et al.* [11] proposed the use of a dense projected pattern with a tuned local operator designed for finding the relationship between blur and depth. Moreno *et al.* [12] proposed the use of an evenly spaced point pattern with defocus to approximate depth in the context of automatic image refocusing. Furthermore, these methods use high density projection patterns which require either a custom projector or more specialized calibrated hardware. Recently, an alternative approach was proposed by Ma *et al.* [2, 3, 4] that leveraged the active projection of quasi-random point projection patterns, which shows considerable promise as it does not require custom projectors or specialized calibration hardware, and thus can enable low-cost, compact depth-sensing systems.

One area that was not well explored in previous work [2, 3, 4] that can be very promising is the use of multispectral quasi-random point projection patterns, which are leveraged in the method designed and implemented in this paper. The use of multiple wavelengths that can be separated when captured using a conventional RGB camera has the potential to increase the spatial resolution of depth measurements made while retaining the simplicity and low complexity of the approach.

3 Method

The depth inference approach via multispectral active depth from defocus and deep learning can be summarized as follows. First, a multispectral quasi-random point pattern is projected onto the scene, which is then captured by a RGB camera. The camera's focus is fixed such that the degree of focus of each point in the projected point pattern as it appears to the camera is dependent on the depth of the surface. Second, the projected pattern as captured by the camera is then passed into an ensemble of deep convolutional neural networks, with each network responsible for estimating the depth of a projected point at a different spectral wavelength. As such, the ensemble of deep convolutional neural networks produces sparse depth measurements at different wavelengths. The final depth map is reconstructed via triangular-based interpolation based on the sparse depth measurements. A one-time calibration step is required to learn the ensemble of deep convolution neural networks.

3.1 Multispectral Quasi-random Point Projection

The core concept underlying the depth inference method is shown in Fig 1. The scene is illuminated by a quasi-random projection pattern consisting of numerous one-pixel points in blue and red wavelengths, and then captured by a RGB camera. When out-of-focus, the projected point will appear blurred, with the degree of blurriness correlated with the depth of the scene at that point. Although both blue and red projection point share the same one-pixel

point structure, the blur effects as captured by the camera can be drastically different. Therefore, given interspersing points at different wavelengths in a quasi-random manner within an active projection, one can achieve higher spatial resolution in the reconstructed depth map.

3.2 Ensemble of Deep Neural Networks for Depth Inference

The purpose of the ensemble of deep convolutional neural networks is to learn and extract intrinsic features to effectively characterize the blurriness of point patterns at different depths at different wavelengths. In the ensemble, each deep convolutional neural network is responsible for performing depth inference at a particular wavelength.

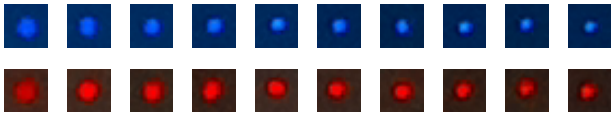

Fig. 1: Projected points on a vertical surface at various distances (left to right: 36cm to 45cm) away from the setup, as captured by a Raspberry Pi camera. The focus for both camera and projector are fixed at 50cm.

To train each deep convolutional neural network in the ensemble, a point pattern is projected onto a vertical surface placed at known distances away from the projector-camera setup. The projected points are extracted from the acquired images and a 20x20 image patch of pixels is formed at each point location and labelled accordingly. For each depth level, a total of four quasi-random point patterns are projected and captured to train and validate the networks. The four point patterns consist of the actual quasi-random point pattern and three one-pixel-shifted versions (horizontal, vertical, and diagonal) of the actual pattern which closely resembles the blurriness of the original pattern.

Point patterns captured from the projection of the three shifted versions of the quasi-random point pattern are used to train the convolutional deep neural networks in the ensemble, and the actual quasi-random pattern is used for testing. There are 3883 points in the original quasi-random point pattern. The three shifted versions of the pattern result in a total of 11,649 20x20 images for each depth label.

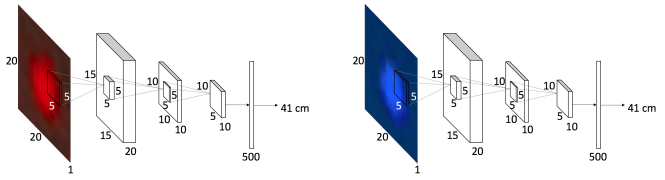

Fig. 2: Architecture of the proposed ensemble of convolutional neural networks for depth inference using red projection pattern(left) and blue projection pattern (right)

The training procedure is visualized in Fig 2. Each deep convolution neural network in the ensemble contains three convolution layers and a fully-connected layer. The use of Rectified Linear Unit layer provides non-linearity at the end of every convolution and fully connected layer. Each network, responsible for a different spectral wavelength, takes the pixel-intensity values from the 20x20 image patch as input and predicts the depth label corresponding to the image patch. The first convolution layer filters the 20x20x1 input image patch with 20 kernels of size 5x5x1. The second convolution layer takes the output of the first convolution layer and filters it with 10 kernels of size 5x5x20. The last convolution layer 5 kernels of size 5x5x10 connected to the output of the second layer.

3.3 Depth Inference Pipeline

With the ensemble of deep convolutional neural networks, the depth of the scene can be estimated. To this end, the depth recovery method can be divided into 3 main stages outlined in Fig 3 and described as follows.

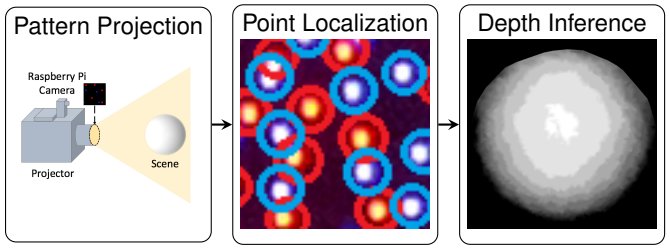

Fig. 3: Illustration of the proposed depth inference pipeline. The scene is actively illuminated with a multispectral quasi-random point pattern and a RGB camera is used to capture images of the projected pattern. The ensemble of deep convolutional neural networks then analyses the captured image and predict depth at each point in the projected pattern. With depth measurements at all locations predicted using the ensemble, triangulation-based interpolation is performed to generated the final depth map.

Stage 1: Multispectral Active Quasi-random Point Projection: A multispectral quasi-random point pattern is projected onto the scene. Poisson-disc sampling (PDS) method was utilized to generate the quasi-random point pattern such that the random points are tightly packed together, but no closer than a specified minimum distance [13]. Given projector resolution $[x,y]$, the PDS algorithm $\phi(\cdot)$ can be expressed as:

$$P = \phi(x,y,\rho,d) \tag{1}$$

where ρ is the desired pattern density, d is the minimum distance between points and P is the quasi-random point map. Compared to other random sampling methods such as Sobol sequence and Halton sequence [14], PDS method significantly reduces the chances of having overlaps between blurred projected points, which would result in erroneous depth recovery. To generate the multispectral point pattern, PDS is performed once for each wavelength and the results are concatenated into a single projection pattern.

Stage 2: Point Localization: After the projected point pattern has been captured by the camera, projection points corresponding to the same wavelength can be effectively separated by taking the single channel measurements from the camera. We use Otsu's method to obtain a binary map consisting of regions of the projected points [15]. The centroid of each region is computed and a 20x20 image patch is formed at each point location.

Stage 3: Depth Inference and Depth Image Reconstruction: After identifying the projected point in the acquired scene, the ensemble of deep convolutional neural networks can then be used to predict the depth corresponding to that projected point. By performing this on all projected points in the quasi-random point projection pattern, the sparse depth estimation can be obtained. With depth measurements at all detected locations, triangulation-based linear interpolation is performed to reconstruct the final depth map.

4 Results

In this section, the efficacy of the depth inference method is demonstrated on test scenes. The main goal of this current realization is to build a compact and portable system to obtain depth information of the scene. For this purpose, the scene is imaged using a Raspberry Pi camera and the multispectral quasi-random point pattern is projected using a BENQ MH630 digital projector.

To investigate the performance, depth inference was performed on two different scenes processing different types of structural details: smooth 3D-printed hemisphere, and complex human hand. For comparison purposes, we compare with two variant of a published method in [3], one for each tested spectral wavelength.

In Fig 4, we illustrate the difference between depth maps generated using [3] and that generated using the proposed multispectral method. For the hemisphere shape, the depth maps produced using the compared method fail to accurately distinguish measurement data from first two depth labels, whereas the multispectral approach produces a smoother surface in the region. Similar improvements can be seen in the hand depth map, where the proposed method produced a significantly improved depth map with clearer depth discrimination in the gap between middle finger and