Received May 5, 2013, accepted May 23, 2013, date of publication July 3, 2013, date of current version July 22, 2013.

Digital Object Identifier 10.1109/ACCESS.2013.2271860

Photogrammetric Bundle Adjustment With Self-Calibration of the PrimeSense 3D Camera Technology: Microsoft Kinect

JACKY C. K. CHOW AND DEREK D. LICHTI
Department of Geomatics Engineering, Schulich School of Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
Corresponding author: J. C. K. Chow (jckchow@ucalgary.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, Alberta Innovates, and Killam Trust.

ABSTRACT The Kinect system is arguably the most popular 3-D camera technology currently on the market. Its application domain is vast and has been deployed in scenarios where accurate geometric measurements are needed. Regarding the PrimeSense technology, a limited amount of work has been devoted to calibrating the Kinect, especially the depth data. The Kinect is, however, inevitably prone to distortions, as independently confirmed by numerous users. An effective method for improving the quality of the Kinect system is by modeling the sensor's systematic errors using bundle adjustment. In this paper, a method for modeling the intrinsic and extrinsic parameters of the infrared and colour cameras, and more importantly the distortions in the depth image, is presented. Through an integrated marker- and feature-based self-calibration, two Kinects were calibrated. A novel approach for modeling the depth systematic errors as a function of lens distortion and relative orientation parameters is shown to be effective. The results show improvements in geometric accuracy up to 53% compared with uncalibrated point clouds captured using the popular software RGBDemo. Systematic depth discontinuities were also reduced and in the check-plane analysis the noise of the Kinect point cloud was reduced by 17%.

INDEX TERMS Kinect, camera calibration, quality assurance, quantization, 3D/stereo scene analysis.

I. INTRODUCTION

The Microsoft Kinect has unarguably made an impact in many scientific disciplines (e.g. computer vision, photogrammetry, and robotics) since its first release in November 2010. Although it began as a controller for the Xbox 360 video game console, it was one of the first low-cost and robust off-the-shelf 3D cameras on the market and thus its audience expanded quickly. One of the first uses adopted outside of gaming was in surgery, where a surgeon could scroll through medical images on a computer screen risk-free – simply with the wave of a hand – at Sunnybrook Hospital in Toronto (Canada) [1].

The Kinect is popular for both industrial and research applications. Commercial solutions based on the Kinect are readily available around the world. For example, iPiSoft offers a markerless optical motion capture system based on the Kinect; ReconstructMe, Manctl, and 4DDynamics turn the Kinect into a handheld scanner for 3D object reconstruction; Faceshift uses it to capture facial motions and expressions; and Fitnect uses it to build a virtual dressing room. In the area of research it has been tested as an aiding device for people who are visually impaired [2], for interactive teaching in classrooms [3], as a portable indoor mapper strapped to humans designed for first responders [4], and for motion capture of hands [5], to name a few.

The wide adoption of the Kinect system has resulted in more than 24 million units being sold as of February 2013. Most users of these units assume their Kinect is well-calibrated and is suitable for a wide-range of applications out-of-the-box. However, [6] tested multiple PrimeSense units and found inaccuracies of up to 1.5 cm. Although high precision of an individual unit is reported (suggesting that temporal averaging is not necessary), variations in accuracy of up to 2 cm between manufactured PrimeSense PS1080 devices were presented. Differences between precision and accuracy can suggest the existence of biases and many researchers have

2169-3536/$31.00 © 2013 IEEE

independently reported systematic distortions in the Kinect point clouds [7]–[9].

To reduce the effect of these systematic errors, numerous efforts have been made in the area of software calibration. One of the first Kinect calibrations was done by [10] in the popular software RGBDemo, where the intrinsic and extrinsic parameters of the infrared (IR) and RGB camera were calibrated based on the OpenCV calibration. Burrus [10] also proposed an algorithm for converting the disparity values to depth measurements; however no calibration routine for deriving these conversion parameters was suggested. This approach can estimate the alignment between the infrared image and the RGB image if libfreenect is used and can improve the alignment if OpenNI is used for data capture.

Similar calibrations for aligning the IR camera with the RGB camera using signalized targets can also be found in [8] where PhotoModeler and Australis were used, and in [11] where PhotoModeler was used. To account for the depth distortions in the point cloud, the disparity values were treated as observations and the baseline distance as well as the distance of the memorized reference pattern, were solved.

Khoshelham and Elberink [11] further added two normalization parameters to their depth calibration explained in [7] for accommodating the quantization of depth values. This approach assumed that there is a zero rotational offset between the infrared camera and projector, and that the depth calibration is independent of the lens distortions in a two-step independent procedure. However, as explained in [12], two-step independent 3D camera calibrations can have its shortcomings. In addition, the Kinect is based on triangulation and hence depth is a function of the image measurements.

Smisek et al. [13] modelled the depth errors in object space by solving for two coefficients of a linear mapping function that minimizes the residuals of a best fit plane. In this case, the depth correction is modelled as a function of distance, which has been shown to be inferior compared to expressing it in image space coordinates [14].

Draelos [15] used depth discontinuities to establish correspondences with the RGB image and used the corners of the planar board for registering the depth image with the RGB image. However, depth discontinuities are unstable and unsuitable for accurate pixel measurements. Furthermore, similar to [13], their depth calibration was independent of the image space.

Based on [16], all of the approaches described thus far do not deliver the optimal set of calibration parameters because 1) the calibrations of the cameras were not performed simultaneously; 2) the error modelling was not performed in image space; or 3) unstable points in the depth map were used. Herrera et al. [16] and [17] presented a method for aligning the RGB image with depth map using the point-on-plane constraint without any depth corrections. As an extension, [14] added a depth distortion model that is dependent on the image space rather than object space in their total system calibration and showed superior performance compared to [13].

Chow et al. [18] has also presented a method for calibrating the Kinect's depth image while simultaneously aligning it with the RGB image using the point-on-plane constraint, but have reported a poor precision in the estimated relative translation and rotation between the depth image and RGB image. To improve the precision they used three orthogonal planes to strengthen the depth and RGB co-registration. Even then, they indicated a lack of constraints to recover the lens distortions of the depth image because the IR images were not used.

Staranowicz and Mariottini [19] compared the approach of [14] and portions of the [17] calibration to a calibration method that uses spheres instead of planes. Their results agreed with [18] and indicated weak recovery of relative orientation parameters between the depth and RGB image when using the point-on-plane approach. However, their work with spherical objects only focused on aligning the depth and RGB images and no calibration model for correcting the depth map was proposed.

With mass production of the Kinect, primarily made for gaming, users attempting to employ this sensor for accuracy-demanding applications such as deformation monitoring and simultaneous localisation and mapping should consider calibrating the Kinect themselves. This paper is an extension of the work in [18] and now includes IR images in the bundle adjustment with self-calibration to improve the alignment between the IR and RGB cameras. It simultaneously calibrates all optical sensors in the Kinect system using a checkerboard pattern and depth measurements measured reliably on the surface of the plane. It is also the first calibration method that explicitly models the rotational offset between the depth image and projector of the Kinect. The depth error resulting from angular misalignments (in particular, from the yaw angle) between stereo pairs has been studied and its effect should not be understated [20]. Although all the data in this paper are captured with the Kinect, it is designed for any devices using the PrimeSense technology (e.g. Asus Xtion PRO, Xtion PRO LIVE and Fotonic P70).

This paper begins by giving a general overview of the Kinect hardware and software for operation in Section 2. Section 3 describes the calibration procedure undertaken in this paper. Section 4 explains the calibration model developed and Section 5 shows the calibration results of two Kinects and discusses the quality of the calibration solution.

II. THE KINECT HARDWARE AND SOFTWARE

The initial Kinect hardware released on November 4, 2010 was designed with the intention of working with the Microsoft Xbox 360 only. The first Kinect for Windows version was not released until February 1, 2012. Compared to the Kinect for Xbox, the Kinect for Windows is not too different aside from offering a closer sensing distance (40 cm instead of 80 cm) and redesigned cabling for easier PC connection. However, in terms of the fundamental depth-sensing principle of the optical sensor at its core, the Kinect for Xbox and the Kinect for Windows are the same. They are also the same as the

3D triangulation-based cameras from Asus and Fotonic, as they are all based on the PrimeSense technology. Differences between these sensors stem from other design specifications; for example, Asus sells their 3D camera with or without a colour (RGB) camera built in.

In general, each one of these systems consists of three optical units: an RGB camera, an IR projector, and an IR camera. The projector emits light in the infrared spectrum and illuminates the scene with a speckle pattern generated from a set of diffraction gratings. Through a 9 by 9, 9 by 7, or 7 by 7 spatial multiplexing window, the pixel showing the highest correspondence among its 64 horizontal neighbours is selected in the infrared image as the corresponding point [21]. Further sub-pixel refinement then gives it a measurement accuracy of approximately 1/8th of a pixel [22].

The optical axes of the projector and IR camera are nominally parallel and are separated by a baseline distance of approximately 7.5 cm. Through photogrammetric triangulation the depth can then be determined. The Kinect stores the disparity value of every pixel at a calibrated distance; therefore a difference between the measured disparity and reference disparity translates into a change in depth [23]. If colour information is desired (for instance in segmentation/classification applications), the RGB camera situated at approximately 2.5 cm from the IR camera can overlay 8-bit 3-channel red, green, blue information over the point cloud.

A standard Kinect has a vertical and horizontal field of view (FOV) of 43° and 57°, respectively. This can be extended by equipping a Kinect with the Nyko Zoom add-on. Although these additional lenses give the Kinect a wider FOV, they will likely increase the magnitude of lens distortion and will require a geometric calibration before they can be used for precise applications.

The raw RGB and IR images, as defined by the Aptina MT9M112 and Micron MT9M001 CMOS sensors respectively, are 1280 pixels by 1024 pixels [24]. Although most APIs allow access to higher resolution images (e.g. SXGA), it comes at the cost of a reduced frame rate. For depth acquisition at 30 Hz using a USB2.0 connection, VGA resolution is usually used due to bandwidth limitations.

For this paper, two Kinect for Xbox sensors were used. All images were captured using the standard VGA resolution to ensure that the IR images are calibrated at the same image resolution as the depth images. Among the various options for operating the Kinect, the Microsoft Kinect SDK was chosen instead of the open source library OpenKinect and the popular OpenNI. Most Kinect calibration work to date has been using OpenNI, as it is conveniently packaged into OpenCV and PCL, but the calibration model in this paper is software-independent and is applicable to Kinect data captured using any driver.

III. CALIBRATION PROCEDURE

Following a two hour warm-up period, depth images, IR images, and RGB images of a checkerboard target were acquired from various positions and orientations. At every exposure, 20 consecutive depth images were captured and averaged to reduce the random noise of the depth measurements and to fill in holes in the depth map. Although [6] and [18] suggested that improvements to range precision through temporal averaging is small (e.g. 1 mm improvement at 3 m distance), likely due to the low depth resolution, it can be done easily.

Since the Kinect cannot capture both the RGB and IR images at the same time, the RGB images and depth images are captured together first. Afterwards, the projector is covered and the IR image of the scene illuminated by an external light source is captured, which is an approach similar to [25]. This ensures good contrast in the IR image in the absence of disturbance from the projector. In the current Microsoft Kinect SDK 1.7 the projector can be switched off, making this step less cumbersome.

The observations in the adjustment can be categorized into three groups: image coordinates in the RGB images (x^{RGB}, y^{RGB}), image coordinates in the IR images (x^{IR}, y^{IR}), and image coordinates as seen by the projector (x^{PRO}, y^{PRO}), which were derived from the depth values retrieved from the SDK. The image coordinates from both cameras were obtained by measuring the corners of a checkerboard pattern using the MATLAB Camera Calibration Toolbox. The depth measurements were made by selecting a randomly distributed set of pixels in the depth image that belonged to the same plane as the checkerboard pattern. For pre-adjustment screening, a plane was fitted to the point clouds derived from the depth images and points were removed using Baarda's data snooping.

IV. MATHEMATICAL MODEL

The user self-calibration method presented in this paper is based on the pin-hole camera model given in Equation 1. To model departures from collinearity, Brown's model [26] for radial lens and decentring lens distortion is augmented with a model for in-plane distortions (i.e. affinity and shear), as shown in Equation 2. Equations 1 and 2 together form the standard photogrammetric bundle adjustment with self-calibration model and are the basis of our proposed calibration method [27].

$$\begin{aligned} x_{ij} &= x_p - c\frac{m_{11}\left(X_i - X_{oj}\right) + m_{12}\left(Y_i - Y_{oj}\right) + m_{13}\left(Z_i - Z_{oj}\right)}{m_{31}\left(X_i - X_{oj}\right) + m_{32}\left(Y_i - Y_{oj}\right) + m_{33}\left(Z_i - Z_{oj}\right)} \\ &\quad + \Delta x \\ y_{ij} &= y_p - c\frac{m_{21}\left(X_i - X_{oj}\right) + m_{22}\left(Y_i - Y_{oj}\right) + m_{23}\left(Z_i - Z_{oj}\right)}{m_{31}\left(X_i - X_{oj}\right) + m_{32}\left(Y_i - Y_{oj}\right) + m_{33}\left(Z_i - Z_{oj}\right)} \\ &\quad + \Delta y \end{aligned} \tag{1}$$

where x_{ij} and y_{ij} are the image coordinates of point i in image j; x_p and y_p are the principal point offsets; c is the principal distance; X_i, Y_i and Z_i are the object space coordinates of point i; X_{oj}, Y_{oj} and Z_{oj} are the position of image j in object space; $m_{11} \ldots m_{33}$ are the elements of the rotation matrix defining the orientation of image j and expressed using the Cardan angle sequence (ω_j, ϕ_j and κ_j); Δx and Δy are the correction terms of additional calibration parameters.

$$\begin{aligned}\Delta x &= x'_{ij}\left(k_1 r_{ij}^2 + k_2 r_{ij}^4 + k_3 r_{ij}^6\right) + p_1\left(r_{ij}^2 + 2x'^2_{ij}\right)\\ &\quad +2p_2 x'_{ij}y'_{ij} + a_1 x'_{ij} + a_2 y'_{ij} \qquad (2)\\ \Delta y &= y'_{ij}\left(k_1 r_{ij}^2 + k_2 r_{ij}^4 + k_3 r_{ij}^6\right) + p_2\left(r_{ij}^2 + 2y'^2_{ij}\right)\\ &\quad +2p_1 x'_{ij}y'_{ij}\end{aligned}$$

where x'_{ij} and y'_{ij} are the image coordinates of point i in image j after correcting for the principal point offset; r_{ij} is the radial distance of point i in image j relative to the principal point; k_1, k_2 and k_3 are the radial lens distortion coefficients; p_1 and p_2 are the decentring lens distortion coefficients; a_1 and a_2 describe the in-plane affinity and shear distortions, respectively.

In this conventional form, the object space target coordinates $\{X_i, Y_i \text{ and } Z_i\}$; interior orientation parameters (IOPs) $\{x_p, y_p \text{ and } c\}$; additional parameters (APs) $\{k_1, k_2, k_3, p_1, p_2, a_1, a_2\}$; and exterior orientation parameters (EOPs) $\{\omega_j, \phi_j, \kappa_j, X_{oj}, Y_{oj}, Z_{oj}\}$ of both the IR and RGB cameras can already be calibrated simultaneously.

As shown in [28] and [29] the calibration of stereo cameras can be improved by constraining the six relative orientation parameters (ROPs) between the stereo pair to be the same at every exposure. Success of ROP constraints has also been demonstrated for systems with three cameras [30] and more [31]. In the least-squares adjustment, this constraint can be realized by adding equations/observations [32], [33] or be integrated directly into the collinearity equations [29], [34]. It was further explained in [34] and [35] that expressing this in the functional model rather than as a constraint equation can reduce the computational load and allow the rotations, translations, and their corresponding standard deviations to be estimated directly.

The optical sensors in the Kinect are rigidly mounted together on a metallic frame. Based on our literature review, there are no reasons to believe that the relative positions and orientations between the internal optical sensors change significantly when being handled with care over a short period of time [6]. Therefore a modified collinearity equation shown in Equation 3 is used for self-calibration instead, where the ROPs are introduced.

The four relevant right-handed coordinate systems (IR = infrared camera, RGB = colour camera, PRO = projector, and OBJ = object space) are illustrated in Figure 1. The notation adopted in this paper is as follows: a superscript of the frame alone means the quantity is expressed in that particular coordinate system (e.g. $[p_{ij}]^{RGB}$ is a vector observed in the RGB-frame); when both subscript and superscript of a coordinate system exist, it represents from *subscript-frame* to *superscript-frame* (e.g. $[R_j]_{OBJ}^{IR}$ is a matrix defining rotations from OBJ-frame to IR-frame).

$$\begin{aligned}&[p_{ij}]^{RGB} - \frac{1}{\mu_{ij}^{RGB}}\Delta R_{IR}^{RGB}\,[R_j]_{OBJ}^{IR}\\ &\qquad\left([O_i]^{OBJ} - [T_j]_{OBJ}^{IR} - [R_j]_{IR}^{OBJ}\, b_{IR}^{RGB}\right) = 0 \qquad (3)\end{aligned}$$

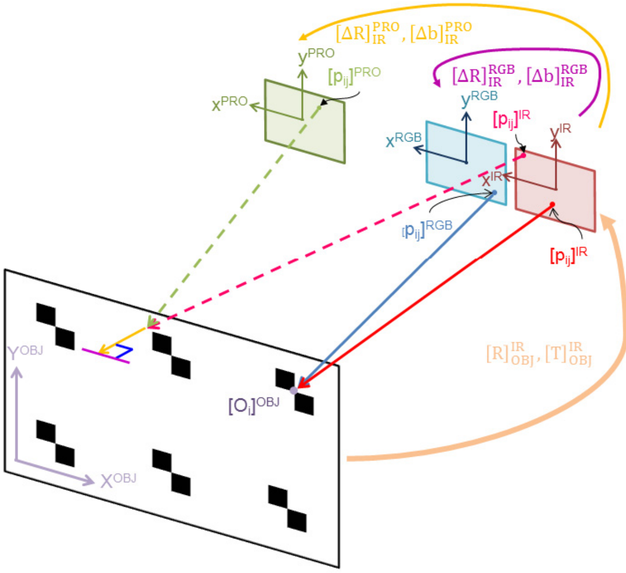


FIGURE 1. Definition of the coordinate systems for the calibration.

where $p_{ij} = \begin{bmatrix} x'_{ij} - \Delta x & y'_{ij} - \Delta y & c \end{bmatrix}^T$ is the corrected image coordinate vector for point i in image j ; μ_{ij} is the unique scale factor for point i in image j; ΔR is the relative rotation matrix defined by rotation angles about the primary ($\Delta\omega$), secondary ($\Delta\phi$) and tertiary axis ($\Delta\kappa$); R_j is the rotation matrix of image j defined by rotations about the primary (ω_j), secondary (ϕ_j) and tertiary axis (κ_j); O_i is the object space coordinates, $[X_i \quad Y_i \quad Z_i]^T$; T_j is the translation vector of image j, $\begin{bmatrix} X_{oj} & Y_{oj} & Z_{oj} \end{bmatrix}^T$; b is the relative translation vector, $\begin{bmatrix} b_x & b_y & b_z \end{bmatrix}^T$.

The above model is sufficient to model the systematic errors in both the IR and RGB cameras; however, it does not characterize the depth measurements of the Kinect. The depth information is determined by triangulation from the IR camera and projector pair. Therefore, it is modelled as a function of the IOPs and APs of both sensors and the six ROPs defining the stereo pair.

As shown by [21] there is a null band in the depth images corresponding to the use of a 9 by 9 or 9 by 7 correlation window. Only after correcting for this offset can the IR images and depth images share the same EOPs, IOPs and APs. A similar correction was done in [11], [13] to align the depth map with the IR image. With knowledge about the intrinsics of the IR camera, the object space coordinates of every pixel in the depth map can be determined using Equation 4.

$$\begin{aligned}X_i^{IR} &= -\frac{Z_i^{IR}}{c^{IR}}\left(x_i^{IR} - x_p^{IR} - \Delta x^{IR}\right)\\ Y_i^{IR} &= -\frac{Z_i^{IR}}{c^{IR}}\left(y_i^{IR} - y_p^{IR} - \Delta y^{IR}\right) \qquad (4)\end{aligned}$$

By knowing the ROPs between the IR camera and projector, the object space coordinates can be back-projected into the image space of the projector. As in most camera-projector calibrations [36], [37] the projector can be treated like a camera; however in this case we do not know the structure of the projected pattern and thus, the IOPs and APs of the projector cannot be recovered reliably. To complicate the problem, the extrinsic parameters between the projector and IR camera are initially unknown. Hence the APs, IOPs, and ROPs of the projector need to be solved iteratively through forward and backward projections due to the non-linear nature of the collinearity equations.

With the image measurements from the RGB camera, the IR camera, and the projector, a bundle adjustment can be performed. To strengthen the calibration, the point-on-plane constraint has been included in the bundle adjustment, which minimizes the residuals orthogonal to the plane [38]. This is necessary for calibrating the depth map, as checkerboard patterns cannot be seen – only geometric features can be identified. Although this restricts the calibration field to a planar 2D target field rather than a 3D volume, a planar target field is portable and is practical for on-site calibration. The drawback of a 2D calibration field (i.e. projective compensation) can be mitigated by imaging a planar checkerboard pattern with converging geometry from various positions and perspectives.

The calibration model for the IR camera and RGB camera with relative translational and rotational constraints was shown in Equation 3. Likewise, the functional model for the depth-projector pair is given in Equation 5. The plane constraint is expressed by the scale factor term μ_{ij} which may be solved by substituting Equation 3 into Equation 6. This final calibration model minimizes the discrepancy between conjugate light rays while constraining them to lie on the best fit plane by solving for the EOPs, ROPs, IOPs, APs, and object space quantities simultaneously.

$$\mu_{\mathrm{ij}}^{\mathrm{PRO}} R_{\mathrm{IR}}^{\mathrm{OBJ}} \mathrm{R}_{\mathrm{PRO}}^{\mathrm{IR}} \left[\mathrm{p_{ij}}\right]^{\mathrm{PRO}} + \left[\mathrm{T_j}\right]_{\mathrm{OBJ}}^{\mathrm{IR}} + \mathrm{R}_{\mathrm{IR}}^{\mathrm{OBJ}} \left[\mathrm{b}\right]_{\mathrm{Pro}}^{\mathrm{IR}} - \left(\mu_{\mathrm{ij}}^{\mathrm{IR}} \mathrm{R}_{\mathrm{IR}}^{\mathrm{OBJ}} \left[\mathrm{p_{ij}}\right]^{\mathrm{IR}} + \left[\mathrm{T_j}\right]_{\mathrm{OBJ}}^{\mathrm{IR}}\right) = 0 \qquad (5)$$

$$\left[\mathrm{a_k} \quad \mathrm{b_k} \quad \mathrm{c_k}\right] \left[\mathrm{O_i}\right]^{\mathrm{OBJ}} - \mathrm{d_k} = 0 \qquad (6)$$

where $\mathrm{a_k}$, $\mathrm{b_k}$, and $\mathrm{c_k}$ are the direction cosines of the normal vector of the best-fit plane; $\mathrm{d_k}$ is the orthogonal distance from the origin to the plane.

Unlike in the MATLAB Camera Calibration Toolbox [39] where the datum is defined by assuming all the object space coordinates are fixed, inner constraints are applied to the object space coordinates, the plane parameters and the EOPs [40]. This is done to improve the overall estimation precision and, most importantly, to prevent possible object space coordinate errors from propagating into the IOPs, APs, and/or ROPs which would result in a biased calibration.

The collinearity and coplanarity equations are highly nonlinear so the Gauss-Helmert least-square model has been chosen for minimizing the summation of the weighted residuals [41]. Baarda's data snooping with a 5% level of significance was used to minimize the possibility of outliers in the adjustment as the least-squares method is known to be highly sensitive to erroneous observations. Iterative variance component estimation has also been adopted for re-weighting the various observation groups (i.e. IR image, depth image, projector, and RGB image) as part of the adjustment [42].

The observations were assumed to be additive zero-mean Gaussian distributed errors uncorrelated with each other [22], and this has shown to be capable of describing the depth noise of the Kinect as a function of distance up to 10 m [18]. The effect of quantization of the disparity measurements is included in the stochastic model [43] and is given in Equation 7. The Kinect's disparities are normalized and quantized for streaming as 11-bit integers with the first bit indicating whether a depth measurement is valid or not, therefore further reducing their range to 1024 levels [11]. As suggested in [22], if the disparity values range between 2 and 88 pixels and are measured with a precision of 1/8 of a pixel ($\sigma_{\mathrm{Disparity}}$), then by using the nominal focal length and baseline distance the effect of quantization step (q) on the depth reconstruction precision (σ_{Z}) as determined by Monte Carlo Simulation (1000 simulations per level) is given in Figure 2. This shows that the quantization has a relatively small effect on the depth reconstruction accuracy, which agrees with the findings reported in [11].

$$E\left\{\text{Total Disparity}^2\right\} = \sigma_{\mathrm{Disparity}}^2 + \frac{\mathrm{q}^2}{12} \qquad (7)$$

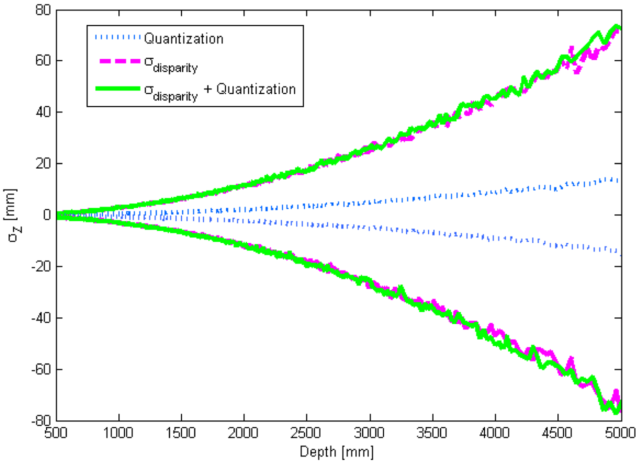


FIGURE 2. Effect of the stochastic model on the accuracy of depth reconstruction. The dashed line (magenta) shows the uncertainty of the reconstructed depth if the disparity measures are continuous. The dotted line (blue) is the effect of quantization. The solid line (green) shows the result of quantized disparity values.

This calibration method follows the three requirements for a 3D camera calibration laid out by [16]. It is "accurate" because the object space target coordinates, plane parameters, EOPs, ROPs, IOPs and APs of all 3 optical sensors are estimated simultaneously in the same bundle adjustment; it is "practical" in that only a single planar checkerboard target is required; and it is "widely applicable" because depth measurements lying in the bounds of the plane are used instead of depth discontinuities.

TABLE 1. Intrinsic parameters recovered from calibration.

	Kinect1		Kinect2	
	IR camera	RGB camera	IR camera	RGB camera
x_p [μm ± μm]	-30.2 ± 2.7	52.1 ± 2.1	-100 ± 1.1	8.5 ± 2.1
y_p [μm ± μm]	-0.4 ± 2.5	-175.9 ± 2.3	-56.8 ± 0.1	-162 ± 1.9
c [mm ± mm]	6.045 ± 0.003	2.896 ± 0.002	6.136 ± 0.001	3.017 ± 0.003
k_1 [1/mm^2 ± 1/mm^2]	-3.9e-3 ±1.8e-4	2.3e-2 ± 1.5e-3	-2.5e-3 ± 8.5e-5	3.1e-2 ± 1.6e-3
k_2 [1/mm^4 ± 1/mm^4]	3.8e-4 ± 3.3e-5	-1.0e-2 ± 1.1e-3	3.6e-4 ± 1.8e-5	-1.2e-2 ± 1.3e-3
k_3 [1/mm^6 ± 1/mm^6]	-1.2e-5 ± 1.8e-6	1.5e-3 ± 2.1e-4	-1.5e-5 ± 1.1e-6	1.6e-3 ± 2.9e-4

TABLE 2. Relative orientation parameters recovered from calibration.

	Kinect1		Kinect2	
	IR-PRO	IR-RGB	IR-PRO	IR-RGB
$\Delta\omega$ [arcsec ± arcsec]	585 ± 81	431 ± 149	42 ± 23	-2690 ± 132
$\Delta\phi$ [arcsec ± arcsec]	844 ± 75	-1970 ± 111	-12 ± 28	-975 ± 121
$\Delta\kappa$ [arcsec ± arcsec]	152 ± 13	-2185 ± 22.7	17 ± 5	-1858 ± 24
b_x [mm ± mm]	76.5 ± 0.1	25.6 ± 0.1	74.1 ± 0.1	27.0 ± 0.3
b_y [mm ± mm]	-0.1 ± 0.1	0.7 ± 0.1	0 ± 0.1	2.2 ± 0.2
b_z [mm ± mm]	-0.9 ± 0.3	0.7 ± 0.3	0.4 ± 0.2	15.4 ± 0.8

Unlike [11], who assumed the calculated depth is independent of the lens distortion, we assume planimetric and depth coordinates are a function of the IR camera's APs. But as in [11] the proposed depth calibration is a function of the IR camera's principal distance and baseline distance. Parameters that account for misalignment between the IR camera's axes and the projector's axes are also included as the Cardan angle sequence.

Similar to [14], every depth pixel has a different calibration coefficient expressed in image coordinate units. However, the number of unknowns being solved is significantly lower in this case because they can be conveniently expressed by the APs. In addition, the IR images are used directly for the mutual de-correlation of parameters rather than using an external high-resolution camera since [14] has demonstrated that the improvement is small and external cameras can complicate the calibration procedure.

In summary, the unknown parameters in the adjustment are the principal point offset, principal distance, and lens distortion parameters of the IR and RGB cameras (IOP^{IR}, AP^{IR}, IOP^{RGB}, AP^{RGB}), the rotational and translational parameters of the projector relative to the IR camera (ROP$^{PRO}_{IR}$) and the RGB camera relative to the IR camera (ROP$^{RGB}_{IR}$), the IR camera orientation and position relative to the object space (EOP$^{IR}_{OBJ}$), the object space target coordinates, and the plane parameters.

Beginning with the initial approximations of zero principal point offset, APs and rotational offsets, a 2.9 mm principal distance and a 7.5 cm positional offset in the x-direction between the IR camera and projector, the depth observations (D^{IR}) were converted into image coordinate measurements of the projector (x^{PRO}, y^{PRO}) by back-projection (Equation 3). The linearized least-squares optimization is then carried out using the math model presented in this section. After bundle adjustment, with the current best estimate of the unknown parameters and D^{IR}, the point cloud is back-projected into the projector again and with the updated x^{PRO} and y^{PRO} the bundle adjustment is repeated until convergence.

V. RESULTS AND DISCUSSION

A. SELF-CALIBRATION RESULTS

To evaluate the proposed method, two Kinect for Xbox sensors were calibrated following the above procedure. One planar target with 24 signalized targets was observed by one Kinect (hereafter "Kinect1") from 11 different poses and 48 planar signalized targets were imaged by another Kinect (hereafter "Kinect2") from 10 different stations (Figure 3). The number of observations in the bundle adjustment is 1972 for Kinect1 and 3152 for Kinect2, with an average redundancy number of 0.93 and 0.94 respectively, yielding a well-controlled network in both cases.

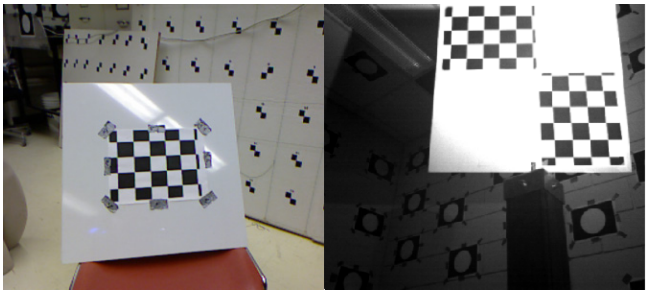


FIGURE 3. (Left) RGB image of the calibration field used for calibrating Kinect1. (Right) IR image of the calibration field used for calibrating Kinect2.

The estimated IOPs and the statistically significant APs for the IR camera and RGB camera are displayed in Table 1. The estimated relative translation and rotation between the three optical sensors are given in Table 2. The initial estimates of 7.5 cm and 2.5 cm for the baseline distances between the IR camera and projector and the IR camera and RGB camera are close to the recovered values. However, detectable rotational offsets between the IR camera and projector were found in one of the Kinects. Compared to the previous work in [18] the standard deviations of ROP$^{RGB}_{IR}$ are significantly improved (by up to 46x for some parameters) with the inclusion of the IR images.

The RMSE of the misclosure of conjugate light rays lying on the best fit plane before and after modelling for the IOPs,

TABLE 3. Quality of plane-fit estimated in the bundle adjustment with self-calibration.

	Kinect1 (mm)			Kinect2 (mm)		
	Before	After	% Improv.	Before	After	% Improv.
RMSE X	5.3	0.1	98%	2.9	0.1	97%
RMSE Y	4.0	0.1	98%	6.3	0.2	97%
RMSE Z	4.2	0.1	98%	3.9	0.1	97%
Std. Dev. of dk	1.1	0.9	18%	0.6	0.5	17%

TABLE 4. Estimated standard deviation of the observation residuals.

	Kinect1 (μm)			Kinect2 (μm)		
	Before	After	% Improv.	Before	After	% Improv.
IR	2.1	2.8	0	2.5	2.6	0%
Depth	28.2	1.3	95%	24.8	0.8	97%
Projector	29.5	1.2	96%	26.0	0.8	97%
RGB	1.6	1.7	0%	2.9	2.8	0%

APs, and ROPs of the depth image and projector pair is given in Table 3. The standard deviation of the plane parameter d_k of the best fit plane as estimated simultaneously by all optical sensors is provided as well, as an indication of the quality of the plane used for this assessment; note that even when the systematic errors of the depth map are untreated, the plane parameters are still well-estimated because of the stereo-pair formed by the calibrated IR and RGB images. Prior to calibration the precisions of the two Kinects differ, but after calibration they are more comparable. An RMSE of up to 6.3 mm was observed before calibration, but after modeling for the systematic errors in the IR camera and projector the RMSE were all less than 0.2 mm.

A plot showing the residuals of the depth map before and after calibration is provided in Figure 4. Before calibration, reprojection errors up to 52 μm (5 pixels) can be perceived but they were reduced to the sub-pixel level after calibration.

Based on the Aptina and Micon sensor's specifications, the pixel size of the IR and RGB cameras is 10.4 μm and 5.6 μm, with a nominal focal length of 6 mm and 2.9 mm, respectively. With a lack of specifications for the projector, it is initially assumed to have the same specifications as the IR camera. Using variance component estimation, the depth image and projector were measuring with a standard deviation of 1/8 of a pixel for Kinect1 and 1/13 of a pixel for Kinect2 after calibration (as seen in Table 4). This is more precise than measurements of the checkerboard pattern made by either camera using the Camera Calibration Toolbox, which delivers approximately 1/4 of a pixel standard deviation for the IR cameras, and 1/3 and 1/2 of a pixel measurement precision for the RGB cameras of Kinect1 and Kinect2 respectively.

With the current calibration model the overall parameter correlation is low: the maximum correlation of the IOPs/APs with the EOPs is 0.26, with the target coordinates is 0.25, and with the plane parameters is 0.14. Some noticeably significant correlations are highlighted in Table 5 (excluding the correlation between the APs, which are already known to exist). Most of the correlation patterns are common between both calibrations, except for the correlations between x_p^{RGB}-$\Delta\phi^{IR}$ and c^{RGB}-b_y^{RGB} which may have higher dependency on the imaging geometry. As the only significant correlations are between the APs and ROPs, these calibration parameters should be transferable to other datasets captured by the same Kinect. To confirm this, additional data were acquired using Kinect1 to assess the accuracy of the calibration.

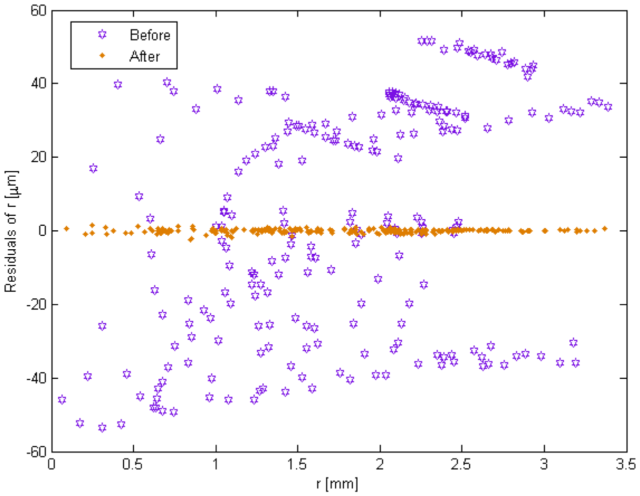


FIGURE 4. Residuals of the depth map as a function of the radial distance from the principal point before and after systematic error modelling.

TABLE 5. Signficant parameter correlations in the bundle adjustment.

	Kinect1	Kinect2
x_p^{IR}-$\Delta\phi^{IR}$	0.97	0.82
y_p^{IR}-$\Delta\omega^{IR}$	0.97	0.94
c^{IR}-ΔZ^{IR}	0.74	0.84
x_p^{IR}-ΔX^{IR}	0.59	0.61
x_p^{RGB}-$\Delta\phi^{IR}$	0.53	0.24
x_p^{RGB}-$\Delta\phi^{RGB}$	0.76	0.88
y_p^{RGB}-$\Delta\omega^{RGB}$	0.85	0.94
c^{RGB}-b_y^{RGB}	0.06	0.47
c^{RGB}-b_z^{RGB}	0.47	0.75

B. EXTERNAL QUALITY ASSESSMENT

To quantify the external accuracy of the Kinect and the benefit of the proposed calibration, a target board located at 1.5–1.8 m away with 20 signalized targets was imaged using an in-house program based on the Microsoft Kinect SDK and with RGBDemo. Spatial distances between the targets were known from surveying using the FARO Focus^{3D} terrestrial laser scanner with a standard deviation of 0.7 mm. By comparing the 10 independent spatial distances measured by the Kinect to those made by the Focus^{3D} the RMSE was 7.8 mm using RGBDemo and 3.7 mm using the calibrated Kinect

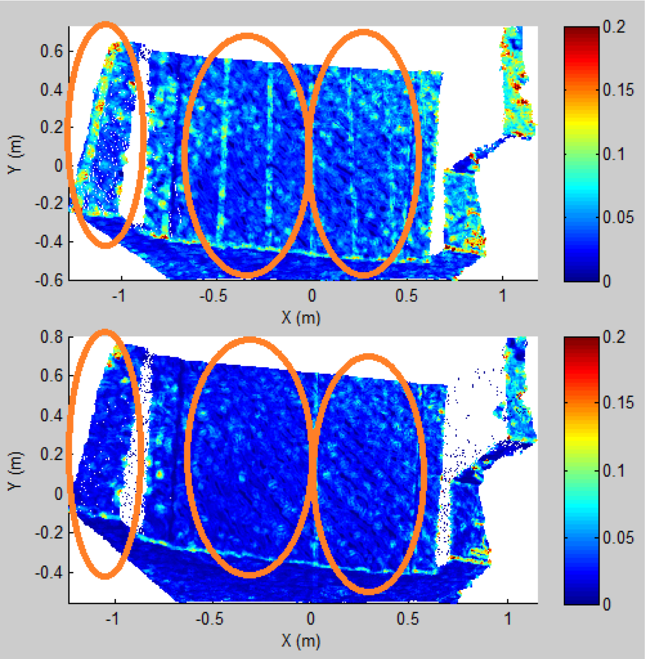


FIGURE 5. (Top) Roughness of point cloud *before* calibration. (Bottom) Roughness of point cloud *after* calibration. The colours indicate the roughness as measured by the normalized smallest eigenvalue.

results; showing a 53% improvement to the accuracy. This accuracy check assesses the quality of all the imaging sensors and not just the IR camera-projector pair alone.

To isolate the assessment of the IR camera-projector quality, the roughness of another scene consisting of a flat target board was computed. The surface roughness was calculated as the normalized smallest eigenvalue in a 30 mm radius neighbourhood. From Figure 5 one can observe that a few systematic abrupt changes in the depth appearing as vertical streaks (highlighted in orange) have been eliminated post calibration. The reason for these stripe artifacts that are parallel to the y-axis of the image coordinate system is unknown as details about the PrimeSense algorithm is still a trade secret. Nonetheless, this artifact was also identified in [8], but was not handled by their calibration scheme. In addition, similar to [14], the presented calibration approach achieves a certain degree of depth smoothing even though lowpass filters were not applied to any of the datasets.

The probability densities of the plane deviations pre- ($\sigma =$ 3.6 mm) and post-calibration ($\sigma = 3.0$ mm) for these data are shown in Figure 6. The noise in depth still follows a Gaussian distribution before calibration, but with a larger standard deviation. In another scene, 20 planes with a 10 cm diameter were extracted. These planes vary in both orientation and position relative to the Kinect, which is important as the residuals from plane fitting are dependent on these parameters. Based on the check plane analysis shown in Figure 7 there is an overall improvement of 17% to the RMSE of the planes estimated using least-squares after calibration.

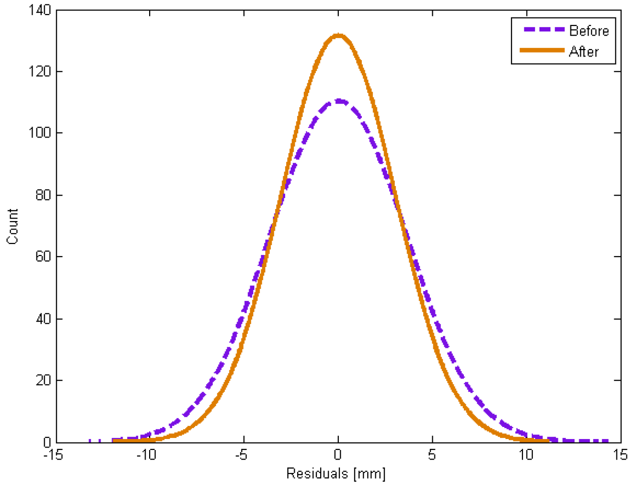


FIGURE 6. Residuals of the plane-fitting before and after calibration.

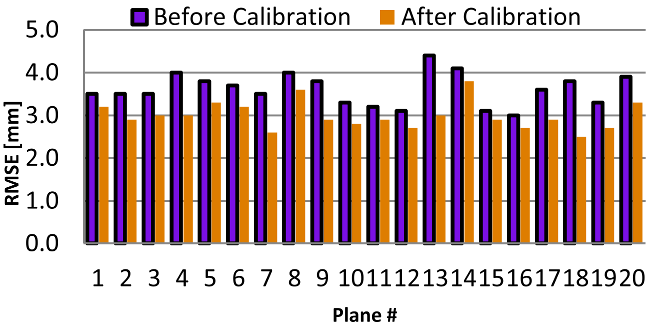


FIGURE 7. RMSE of check planes before and after applying the calibration parameters.

VI. CONCLUSION

A self-calibration method suitable for the Microsoft Kinect was presented and tested. The method solves for the relative translations and rotations between the IR camera, projector, and RGB camera. At the same time, it solves for the intrinsic parameters of both cameras, extrinisic parameters of the IR camera, object space target coordinates, and plane parameters. Geometric constraints have been included in the bundle adjustment to ensure points lie on the best fit plane and the optical sensors are all mounted rigidly on the same platform. The depth calibration is expressed as a function of three rotations, three translations, interior orientation parameters and the lens distortion of the IR camera. The effect of quantized disparity values on depth reconstruction is modelled stochastically in the bundle adjustment despite its small effect on reconstruction accuracy.

In the experimental results, significant rotational offsets up to 0.2 degs between the IR camera and projector have been recovered in the bundle adjustment. Through the inclusion of IR images and various geometric constraints, no significant correlations can be identified between the system calibration parameters and the scene dependent parameters. In the quality control stage, both the precision and accuracy of the Kinect were improved by 17% and 53%, respectively, following the presented calibration method. Furthermore, through qualitative assessment, some visually identifiable systematic artifacts in the Kinect point cloud have been removed.

Future work will study both the short-term and long-term stability of the calibration parameters for these low-cost gaming sensors. Additional features such as lines will be added to the bundle adjustment to improve the precision of the recovered calibration parameters. Recovery of the distortions in the projector could not be performed reliably using the proposed method and will be investigated.

REFERENCES

[1] P. Loriggio, *Toronto Doctors Try Microsoft Kinect in OR*, Toronto, ON, Canada: Globe and Mail, 2011.

[2] M. Zöllner, S. Huber, H.-C. Jetter, and H. Reiterer, "NAVI—A proof of concept of a mobile navigational aid for visually impaired based on the microsoft kinect," in *Proc. 13th IFIP TC* , Lisbon, Portugal, Sep. 2011, pp. 584–587.

[3] M. Johnson and K. Hawick, "Teaching computational science and simulations using interactive depth-of-field technologies," in *Proc. Int. Conf Frontiers Educ., Comput. Sci. Comput. Eng.*, 2012, pp. 1–7.

[4] M. Fallon, H. Johannsson, J. Brookshire, S. Teller, and J. Leonard, "Sensor fusion for flexible human-portable building-scale mapping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Algarve, Portugal, Oct. 2012, pp. 4405–4412.

[5] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3D tracking of hand articulations using kinect," in *Proc. BMVC*, Sep. 2011, p. 101.

[6] J. Boehm, "Natural user interface sensors for human body measurement," *Int. Archive Photogram. Remote Sens.*, 2012, pp. 531–536.

[7] K. Khoshelham, "Accuracy analysis of kinect depth data," in *Proc. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2011, pp. 133–138.

[8] F. Menna, F. Remondino, R. Battisti, and E. Nocerino, "Geometric investigation of a gaming active device," *Proc. SPIE*, vol. 8085, p. 80850G, May 2011.

[9] C. Toth, B. Molnar, A. Zaydak, and D. Grejner-Brzezinska, "Calibrating the MS kinect sensor," in *Proc. ASPRS*, Mar. 2012.

[10] N. Burrus. (2012, Mar. 30). *RGBDemo* [Online]. Available: http://labs.manctl.com/rgbdemo/

[11] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[12] D. Lichti and C. Kim, "A comparison of three geometric self-calibration methods for range cameras," *Remote Sens.*, vol. 3, no. 5, pp. 1014–1028, 2011.

[13] J. Smisek, M. Jancosek, and T. Pajdla, "3D with kinect," in *Proc. IEEE ICCV Workshops*, Barcelona, Spain, Nov. 2011, pp. 1154–1160.

[14] D. Herrera, J. Kannala, and J. Heikkilä, "Joint depth and color camera calibration with distortion correction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2058–2064, Oct. 2012.

[15] M. Draelos, "The Kinect up close: Modifications for short-range depth imaging," M.S. thesis, North Carolina State Univ., Raleigh, NC, USA, 2012.

[16] C. Herrera, J. Kannala, and J. Heikkilä, "Accurate and practical calibration of a depth and color camera pair," in *Proc. 14th Int. Conf. Comput. Anal. Images Patterns*, Seville, Spain, 2011, pp. 437–445.

[17] C. Zhang and Z. Zhang, "Calibration between depth and color sensors for commodity depth cameras," in *Proc. IEEE ICME*, Jul. 2011, pp. 1–6.

[18] J. Chow, K. Ang, D. Lichti, and W. Teskey, "Performance analysis of a low-cost triangulation-based 3D camera: Microsoft Kinect system," in *Proc. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2012, pp. 175–180.

[19] A. Staranowicz and G. Mariottini, "A comparative study of calibration methods for Kinect-style cameras," in *Proc. 5th Int. Conf. Pervas. Technol. Rel. Assistive Environ.*, Jun. 2012, p. 49.

[20] W. Zhao and N. Nandhakumar, "Effects of camera alignment errors on stereoscopic depth estimates," *Pattern Recognit.* vol. 29. no. 12, pp. 2115–2126, 1996.

[21] K. Konolige and P. Mihelich. (2013, Feb. 17). *Kinect Calibration: Technical* [Online]. Available: http://www.ros.org/wiki/kinect_calibration/technical

[22] C. D. Mutto, P. Zanuttigh, and G. Cortelazzo, *Time-of-Flight Cameras and Microsoft Kinect*. New York, NY, USA: Springer-Verlag, 2013.

[23] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli, U.S. Patent 0 018 123, 2010.

[24] (2013, Feb. 19). *Hardware Info* [Online]. Available: http://openkinect.org/wiki/Hardware_info

[25] J. Smisek and T. Pajdla, "3D camera calibration," M.S. thesis, Czech Tech. Univ. Prague, Prague, Czech Republic, 2011.

[26] D. Brown, "Close-range camera calibration," *Photogram. Eng.*, vol. 37, no. 8, pp. 855–866, 1971.

[27] C. Fraser, "Digital camera self-calibration," *ISPRS J. Photogram. Remote Sens.*, vol. 52, no. 4, pp. 149–159, 1997.

[28] G. He, K. Novak, and W. Feng, "Stereo camera system calibration with relative orientation constraints," *Proc. SPIE*, vol. 1820, pp. 2–8, Nov. 1992.

[29] B. King, "Bundle adjustment of constrained stereopairs—Mathematical models," *Geomat. Res. Australasia*, vol. 63, pp. 67–92, Feb. 1995.

[30] A. Tommaselli, M. Galo, J. Marcato, R. Ruy, and R. Lopes, "Registration and fusion of multiple images acquired with medium format cameras," in *Proc. Int. Archives Photogram. Remote Sens.*, 2010, pp. 1–6.

[31] N. El-Sheimy, "A mobile multi-sensor system for GIS applications in urban centers," *Int. Archives Photogram. Remote Sens.*, 1992, pp. 95–100.

[32] J. Lerma, S. Navarro, M. Cabrelles, and A. Seguí, "Camera calibration with baseline distance constraints," *Photogram. Rec.*, vol. 25, no. 130, pp. 140–158, 2010.

[33] A. Tommaselli, M. Moraes, J. Marcato, C. Caldeira, R. Lopes, and M. Galo, "Using relative orientation constraints to produce virtual images from oblique frames," in *Proc. Int. Archives Photogram. Remote Sens.*, 2012, pp. 61–66.

[34] A. Habib, A. Kersting, K. Bang, and J. Rau, "A novel single-step procedure for the calibration of the mounting parameters of a multi-camera terrestrial mobile mapping system," *Archives Photogram., Cartography Remote Sens.*, vol. 22, pp. 173–185, Jan. 2011.

[35] A. Kersting, "Quality assurance of multi-sensor systems," Ph.D. dissertation, Dept. Geomatics Eng., Univ. Calgary, Calgary, AB, Canada, 2011.

[36] M. Trobina, "Error model of a coded-light range sensor," Commun. Technol. Lab. Image Science Group, ETH-Zentrum, Zurich, Tech. Rep., 1995.

[37] G. Falcao, N. Hurtos, and J. Massich, "Plane-based calibration of a projector-camera system," VIBOT, Le Creusot, France, Tech. Rep., 2008.

[38] M. Obaidat and K. Wong, "Geometric calibration of CCD camera using planar object," *J. Survey. Eng.*, vol. 122, no. 3, pp. 97–113, 1996.

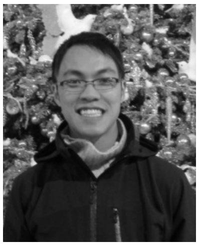
[39] J. Bouguet. (2012, Mar. 30). *Camera Calibration Toolbox for Matlab* [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/index.html

[40] D. Lichti and J. Chow, "Inner constraints for planar features," *Photogram. Rec.*, vol. 28, no. 141, pp. 74–85, 2013.

[41] W. Förstner and B. Wrobel, "Mathematical concepts in photogrammetry," in *Manual of Photogrammetry*, 5th ed. Bethesda, MD, USA: American Society of Photogrammetry and Remote Sensing, 2004, pp. 15–180.

[42] W. Caspary and J. Rüeger, *Concepts of Network and Deformation Analysis*, Kensington, Australia: Univ. New South Wales, 1987.

[43] B. Widrow, I. Kollár, and M. Liu, "Statistical theory of quantization," *IEEE Trans. Intrum. Meas.*, vol. 45, no. 2, pp. 353–361, Apr. 1996.



JACKY C. K. CHOW received the B.Sc. (Hons.) degree and APEGGA Gold Medal in geomatics engineering from the University of Calgary, Calgary, AB, Canada, in 2009. He is currently an iCORE, NSERC, and Killam Scholar. He is currently pursuing the Ph.D. degree in geomatics engineering with the University of Calgary. His current research interests include 3-D imaging, navigation, multisensor integration, and deformation monitoring.

DEREK D. LICHTI received the B.S. (Hons.) degree in survey engineering from Ryerson University, Toronto, ON, Canada, in 1993, and the M.S. and Ph.D. degrees in geomatics engineering from the University of Calgary, Calgary, AB, Canada, in 1996 and 1999, respectively. In 1999, he joined the Curtin University of Technology, Perth, Australia, where he was with the Department of Spatial Sciences in 2007. He joined the Department of Geomatics Engineering, University of Calgary, where he is currently a Professor, in January 2008. His current research interests include developing geomatics engineering solutions for the exploitation of optical and range-imaging sensors, primarily time-of-flight range cameras but also laser scanners and gaming sensors, for the automated creation of accurate, and 3-D models in support of several applications, including the measurement of structures and human motion.