# A Multi-Step Explicit Stereo Camera Calibration Approach to Improve Euclidean Accuracy of Large-Scale 3D Reconstruction

Habib Fathi[1] and Ioannis Brilakis[2]

[1] PhD, CTO, Pointivo LLC., Atlanta, GA, USA; Tel: +1 404-713-3667; Email: habib@pointivo.com

[2] Laing O'Rourke Lecturer of Construction Engineering, Department of Engineering, University of Cambridge, UK; Tel: +44 (0) 1223 332718; Email: ib340@cam.ac.uk

**Corresponding Author:**

Habib Fathi

**Abstract:**

The spatial accuracy of point clouds generated by stereo image-based 3D reconstruction algorithms is very sensitive to the intrinsic and extrinsic camera parameters determined during camera calibration. The existing camera calibration algorithms induce a significant amount of error due to poor estimation accuracies in camera parameters when they are used for large-scale scenes such as mapping civil infrastructure. This leads to higher uncertainties in the location of 3D points, and may result in the failure of the whole reconstruction process. This paper proposes a novel procedure to address this problem. It hypothesizes that a set of multiple calibrations created by videotaping a moving calibration pattern along a specific path can increase overall calibration accuracy. This is achieved by using conventional camera calibration algorithms to perform separate estimations for some predefined distance values. The result, which includes multiple sets of camera parameters, is then used in the Structure from Motion process to improve the Euclidean accuracy of the reconstruction. The proposed method has been tested on infrastructure scenes and experimental analyses indicate more than 25% improvement in the spatial accuracy of 3D points.

**Keywords:**

Camera Calibration; Structure from Motion; Videogrammetry; Bundle Adjustment; Lens Distortion;

## 1. Introduction

3D laser scanning is used in the current practice to capture accurate and detailed spatial data for projects requiring fully automated 3D data retrieval. However, its application in the construction industry is limited due to high costs of the equipment and the necessity for having skilled equipment operators at the field (Klein, et al., 2012). Automatic 3D reconstruction from multiple view imagery or video streams is emerging as an inexpensive alternative to the laser-based systems. Several studies have demonstrated the capabilities of image-based 3D reconstruction approaches that work based on Structure from Motion (SfM) techniques (Furukawa & Ponce, 2010) (Fathi & Brilakis, 2011) (Dai, et al., 2013) (Wu, 2013) (Fathi & Brilakis, 2013) (Golparvar-Fard, et al., 2014). The image-based methods cannot replace the laser-based systems without achieving acceptable levels of geometrical accuracy. Two main issues need to be studied for this purpose: camera calibration and dense multi-view geometry (Strecha, et al., 2008). The scope of this paper is to focus on the first issue as the first step in the stereo-based 3D reconstruction pipeline.

Camera calibration is the process of determining a set of camera parameters that describe the mapping between 3D world coordinates and 2D image coordinates. They are categorized into intrinsic and extrinsic parameters. Intrinsic parameters represent internal geometry and optical characteristics of the lens such as focal length, principal point, and distortion coefficients. Extrinsic parameters represent the camera position and orientation in the 3D world coordinate system. The existing camera calibration methods are divided into two categories: a) explicit calibration (i.e., conventional approach) and b) self- or auto-calibration. For explicit calibration, the parameters are estimated by establishing correspondences between reference points on an object with known 3D dimensions and their projection on the image. On the other hand, self-

calibration automatically calculates the parameters using geometrical constraints in images, but is less accurate than the explicit methods (Furukawa & Ponce, 2009).

Most of the existing image-based 3D reconstruction methods use a single camera as the sensor system. This imposes a constraint on the generated results: using a single camera for image acquisition, the scene can only be reconstructed up to an unknown scale factor (Pollefeys, et al., 2008). This limitation is of great importance especially in infrastructure applications that require spatial data collection in the Euclidean space. The use of a calibrated stereo camera set eliminates this problem, but at the cost of additional steps for camera calibration and more sensitivity of the results to the calibration parameters (Peng, 2011) (Xu, et al., 2012). Comprehensive sensitivity analyses of stereo camera calibration parameters with respect to different factors (e.g., baseline distance, depth of points, etc.) have shown higher uncertainties when the camera distance to the calibration object (hereafter is called depth of calibration) varies in a wider range (Dang, et al., 2009) (Peng, 2011). While testing these theoretical findings through several observations, the authors noticed less uncertainties in estimated camera parameters as well as more accurate 3D reconstruction results for 3D points that have a depth value close to the depth of calibration; hence, the authors hypothesized that the use of multiple sets of calibration parameters rather than a single set could potentially enhance the 3D reconstruction accuracy.

This paper aims to initially demonstrate that the existing calibration procedures could induce significant amount of error if they are used for large-scale 3D reconstruction applications and then test the abovementioned hypothesis. As the main contribution of the paper, a transformational approach is presented for stereo camera calibration and its application in the 3D reconstruction pipeline. This approach does not provide new mathematical relationships for

estimating the necessary parameters; instead, conventional calibration algorithms are used in a multi-step procedure that will result in multiple sets of camera parameters each optimized for a particular depth of calibration. These sets are ultimately used in the bundle adjustment step of the SfM process to achieve optimum X, Y, and Z coordinates for 3D points. This allows maintaining the well-known benefits of the conventional algorithms while improving the Euclidean accuracy of the reconstruction. Results from two case studies show significant improvement in the accuracy of 3D points when using the new approach versus the conventional methods: a reduction of 3.2cm (25%) in terms of the 95% spatial distance error between 3D points in the first experiment and 4.8cm (29%) in the second experiment.

## 2. Background

An ideal camera behavior (i.e., a distortion-free lens) is normally described using a pinhole camera model (Fig. 1). In this model, the camera aperture is defined as a single point and the main assumption is that no lenses are used to focus light. Interested readers are referred to (Hartley & Zisserman, 2003) for mathematical details. In case of non-negligible lens distortion

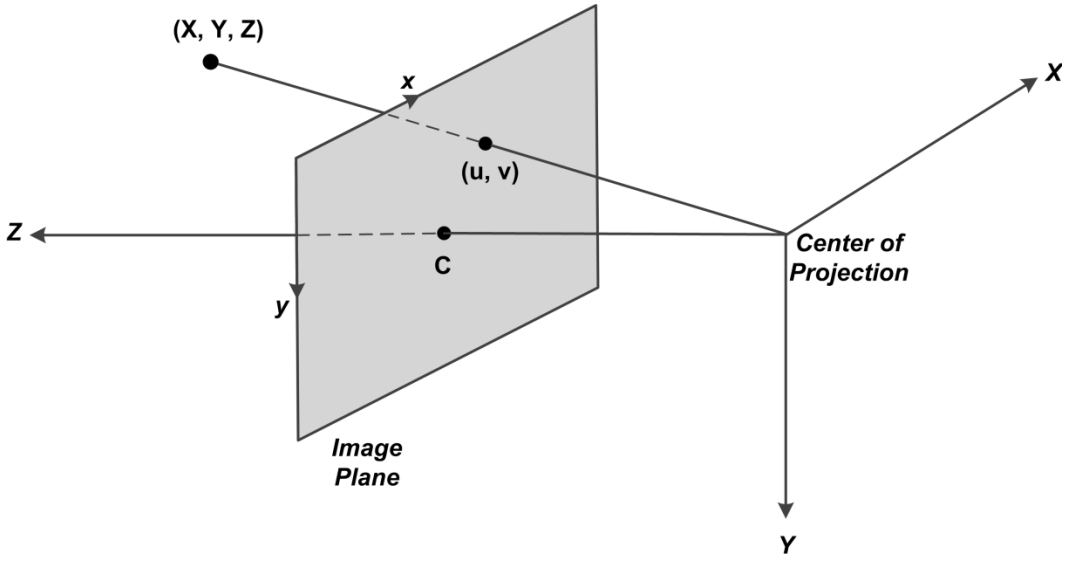



**Figure 1:** A schematic of the pinhole camera model

or far-range 3D reconstruction (i.e., camera distance to the object of interest is more than 10m), pinhole camera model is not accurate enough and more parameters should be taken into account (Ricolfe-Viala & Sanchez-Salmeron, 2010). These parameters are generally used to model the radial and tangential lens distortion through non-linear functions $\delta_u(u,v)$ and $\delta_v(u,v)$ that map the unobservable distortion-free image coordinates $(u,v)$ to image coordinates with distortion $(u_d, v_d)$. The use of these distortion types has been shown to be sufficient for most practical applications (Hartley & Kang, 2005).

In case of monocular 3D reconstruction, the scope of camera calibration narrows down to estimating the focal length $(f_x, f_y)$, principal point $C$, and distortion coefficients. There are two primary approaches to this problem: conventional calibration via scene constraints of objects with precisely known geometry; and self-calibration via SfM (scene geometry plus camera parameters).

(Zhang, 2000) proposed a calibration technique that requires a camera to observe a planar grid-pattern from different orientations. The minimum number of views is two but more views should be used in practice to achieve acceptable results. The redundancy helps to propagate the estimation errors and hence reduce the uncertainty of the estimated parameters. A closed-form solution was presented to model the radial lens distortion which is followed by a non-linear refinement. (Liebowitz, 2001) proved that using a planar object is equivalent to fitting the image of the absolute conic to the image circular points of the imaged plane. Similar approaches were used by others to expand the set of acceptable calibration patterns. The use of circle-like patterns was studied in (Kim, et al., 2005) to take advantage of the fact that imaged circular points are always on the images of planar circles. In another study, (Zhang, 2004) demonstrated that camera

calibration is feasible using only one-dimensional objects (i.e., points aligned on a line), thus eliminating the need for a planar 2D calibration object. However, the method fails if the object is moving freely in the environment. A generic calibration concept was also presented in (Sturm & Ramalingam, 2004) that considers most projection models used in computer vision (e.g., perspective and affine models, optical distortion models, and catadioptric systems), but it is a conceptual model without any quantitative evaluations. More recently, (Wang, et al., 2008) presented a model for lens distortion which used radial distortion parameters plus a transform from the ideal image plane to the real sensor array plane. Although this model has fewer parameters to be calibrated, it has been shown that its performance is very similar to the conventional models. (He & Li, 2008) and (Ricolfe-Viala & Sanchez-Salmeron, 2010) also achieved similar performances by applying different concepts: vanishing points and computing the camera lens distortion isolated from the camera calibration process.

The need for a calibration object has been eliminated through self- or auto-calibration methods. (Furukawa & Ponce, 2009) used a top-down information approach for this purpose (output of a multi-view stereo camera system on scaled-down input images is used to establish feature correspondences). The method requires rough camera parameter estimates which can be acquired from the EXIF tags of still photographs. In another study, (Kim & Kweon, 2009) proposed to use scene constraints in the form of camera constraints which is based on image warping using images of parallelograms. Although these methods have demonstrated promising results in several case studies, they cannot still provide the same level of accuracy that explicit (i.e., calibration object-based) methods have achieved.

Stereo camera calibration is similar to the monocular case with a difference that the relative position between two cameras $(R_0, t_0)$ needs to be found beside the intrinsic and

extrinsic parameters of each camera. The geometric relationship between the left $(R_l, t_l)$ and right $(R_r, t_r)$ cameras can be expressed as follows

| | |
|---|---|
| $R_0 = R_r R_l^{-1} \quad , \quad t_0 = t_r - R_r R_l^{-1} t_l$ | (1) |

A feature-based calibration method for distributed stereo camera networks presented in (Mavrinac, et al., 2010). The method converges, provides pairwise orientations, and scales with network size, but has shown repeatability problems especially in the local interest point detection step. In another study, (Xiao, et al., 2010) designed an accurate stereo camera calibration process for industrial on-site inspection by using a cross-shaped calibration target. The method, however, is only applicable in controlled indoor environments. (Xu, et al., 2012) studied the use of a chessboard-like calibration pattern along with a gradient threshold-based corner extraction method for a stereo vision calibration process. They primarily used the calibration process developed in the Camera Calibration Toolbox, which is based upon methods proposed in (Zhang, 2000) (Sturm & Ramalingam, 2004). The experimental results showed stability and accuracy for a visual system with large baseline (i.e., ~60cm).

The abovementioned methods have been successfully used in close-range 3D reconstruction applications ($Z \leq 2$m) with spatial accuracies that rival laser scanning (Seitz, et al., 2006); however, the same level of accuracy has not been achieved in far-range applications ($Z > 10$m) even using cameras with multi megapixel resolution; (Dai, et al., 2013) shows that errors in the order of ±6-8cm should be expected in such applications.

In a stereo reconstruction problem, in particular, the accuracy of results could be very sensitive to the intrinsic and extrinsic calibration parameters as well as the distance between the

camera and the object of interest (House & Nickels, 2006) (Geiger, et al., 2011). This may be justified by the point that in such a problem, the estimated parameters are kept constant throughout the SfM process and errors can accumulate. (Dang, et al., 2009) have presented a thorough mathematical analysis for sensitivity of stereo 3D reconstruction to erroneous calibration parameters. The result of this study is summarized in Table 1, where $Z$ is the depth of the point to be reconstructed; $b$ is the baseline; $f$ is the focal length in pixels; $C_x$ is the x-coordinate of the camera center; $u$ is the x-coordinate of the point in the image space; $(\tilde{u},\tilde{v})$ are normalized coordinates of the point in the image space; and subscript $L$ denotes the left camera in the stereo rig. From this table, it can be concluded that the sensitivity of the results is the highest for yaw, pitch, and roll. Reconstruction errors also scale linearly with $\Delta b$ and higher tolerances are acceptable in estimating $b$.

**Table 1:** Sensitivity of stereo 3D reconstruction to erroneous calibration parameters (Dang et al., 2009)

| **Error Source** | **Linear Sensitivity of 3D Reconstruction** | **Error Source** | **Linear Sensitivity of 3D Reconstruction** |
|---|---|---|---|
| yaw error $\Delta\Psi_L$ | $\frac{\Delta Z}{\Delta\Psi_L} \approx -\frac{Z^2}{b}\left(1+\tilde{u}_L^2\right)$ | baseline error $\Delta b$ | $\frac{\Delta Z}{\Delta b} \approx -\frac{Z}{b}$ |
| pitch error $\Delta\Phi_L$ | $\frac{\Delta Z}{\Delta\Phi_L} \approx \frac{Z^2}{b}\tilde{u}_L\tilde{v}_L$ | center offset $\Delta C_L$ | $\frac{\Delta Z}{\Delta C_L} \approx -\frac{Z^2}{bf}$ |
| roll error $\Delta\Theta_L$ | $\frac{\Delta Z}{\Delta\Theta_L} \approx \frac{Z^2}{b}\tilde{v}_L$ | focal length error $\Delta f_L$ | $\frac{\Delta Z}{\Delta f_L} \approx \frac{Z^2}{bf^2}\left(u_L - C_x\right)$ |

Existing camera calibration packages such as Camera Calibration Toolbox not only provide the best estimation for each parameter but also calculate the amount of uncertainties in

the given estimation (in terms of a ± range). For example, in the case of calibrating a stereo rig with two 8 megapixel cameras, 12mm focal length, and 28cm baseline, the following output could be achieved: $f_{left} = 3854.01 \pm 17.34$, $C_{left} = (1658.33 \quad 1140.15) \pm (32.21 \quad 34.61)$, $f_{right} = 3845.61 \pm 21.21$, $C_{right} = (1759.26 \quad 1155.42) \pm (35.30 \quad 33.40)$, $T = (-279.30 \quad 0.43 \quad -7.15) \pm (0.76 \quad 0.37 \quad 5.00)$. This range of uncertainty is another source of information that could be used to analyze the sensitivity of the process. An observation in (Peng, 2011) indicates that if the depth of calibration is (more or less) kept constant, the ranges of the uncertainties decrease. On the other hand, if the depth keeps varying in a larger range, higher uncertainties in the estimated parameters could be seen. The following reasons could be listed for such a behavior. First, the projection function, that includes the process to compensate for the lens distortion, is a non-linear function; hence, if the input data covers a broader range of depths, there is a higher probability to be trapped in local optima. Second, the cost function in the optimization process is the reprojection error which is more sensitive to the data in closer depths; hence, the estimated parameters could result in high spatial distance error for data in farther depths.

The problem statement can be summarized as follows. The sensitivity analyses of the stereo 3D reconstruction process have shown a near quadratic relationship between the $Z$ value of a 3D point and the errors in estimated spatial coordinates of the point. This may not be a significant issue in close-range 3D reconstruction as the range for $Z$ is limited ($Z \leq 2$m), but considerably affects the reconstruction accuracy in far-range scenarios ($Z > 10$m). Since this relationship cannot be changed, the only solution to decrease the reconstruction error is to reduce the uncertainties in the estimated camera calibration parameters. This, for example, implies that if the uncertainties were reduced by 50%, the reconstruction accuracy would increase four times.

None of the existing stereo camera calibration procedures address this issue. The research objective of this paper is therefore to enhance the Euclidean accuracy of generated point clouds from stereo 3D reconstruction algorithms using an explicit camera calibration procedure that reduces the level of uncertainties in estimated parameters. The key research question that will be answered is: how can we use the known information about the sensitivity of 3D coordinates of points with respect to calibration parameters and design a camera calibration procedure that is capable of providing higher Euclidean accuracies?

## 3. Solution Hypothesis

Inspired by the outcome of previous studies ( (House & Nickels, 2006) (Strecha, et al., 2008) (Peng, 2011) (Xu, et al., 2012)), the authors performed several observations to assess the amount of uncertainty that may exist in estimating calibration parameters. In these observations, two 5megapixel video cameras with a baseline distance of 30cm as well as fixed focal length lenses with $f = 25mm$ were used to implement two scenarios each for five times: a) depth of calibration varies from 5m to 15m; and b) depth of calibration is fixed at 10m. The reason for repeating each scenario was to study whether they could all result in the same calibration parameters or not. Observations from the first scenario indicated that 5-10% difference can be seen for estimated parameters at each repetition while those differences were 3-4% for the second scenario. Next, the average of the parameters at each scenario were used to calculate 3D coordinates of a set of corresponding feature points in a stereo frame (i.e., a pair of left and right view frames) captured from an object at the distance of ~10m from the camera set. The results for the second scenario showed less noise and spatial distance error.

The initial observations explained above have led the authors to hypothesize that a multi-step stereo camera calibration procedure can enhance the final Euclidean accuracy of 3D points

if the data is properly fused into the reconstruction process. This hypothesis relies upon finding sets of camera parameters each for a particular depth of calibration. These sets of parameters are extracted from the video streams captured from a moving calibration board. As seen before, it is expected to achieve higher reconstruction accuracies if the depth value of a 3D point is close to the depth of calibration.

The following delimitations are listed to clarify the boundaries for the proposed hypothesis in this paper: 1) only a planar calibration object with a chessboard pattern is used and other possible types and shapes of calibration objects are not considered; 2) the method is not applicable for fish-eye lenses; 3) the absolute 3D coordinates of points cannot be measured and hence the accuracy is quantified using the spatial distance between pairs of 3D points; and 4) the effective range of depth values for 3D points is limited to 100 times of the baseline distance due to significant errors that are expected beyond that (Gallup, 2011).

## 4. Methodology

If the depth of calibration is denoted by $D$, a conventional explicit stereo camera calibration procedure is repeated $n$ times for different $D$ values. At each repetition, a different value is selected for $D$ (e.g., $D_i = 10m$) and while it is kept constant, a set of stereo video streams are collected (Fig. 2). During the video recording process, the camera system and the board move in a way that $D_i$ does not change significantly. As a requirement, the calibration board should be videotaped from different angles and the whole pattern should be visible in all video frames. The motion can be arbitrary and does not need to be known, but should not be a pure translation; this constraint is imposed by the conventional camera calibration algorithms. The best strategy could be keeping the camera set in a fixed location and instead moving the calibration board in

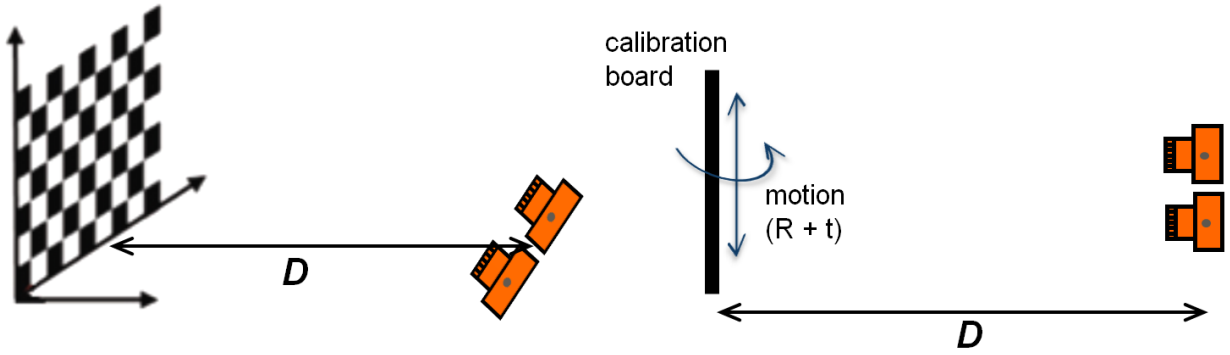



**Figure 2:** Data collection for camera calibration at *D* (left: side view; right: top-down view)

different directions and angles. The directions could be up, down, left, and right. It is also recommended that the board is tilted forward and backward while having up to 45 degrees lateral rotations. It is also necessary to mention that while videotaping, it is preferred to limit the movement such that the calibration board appears at different areas of video frames. It is known that if the calibration pattern only appears at the central part of video frames, the estimations will behave poorly at peripheral areas (Zhang, 2000).

The collected data is then used as the input in a conventional calibration algorithm to find the required parameters. The result corresponding to $D_i$ is saved and the process is repeated for the next $D$ in the sequence. The outcome includes multiple sets of calibration parameters $\{P_i \mid i = 1,...,n\}$, each corresponding to a specific $D$. Once the required calibration information is acquired, $P_i$ is hardcoded into the 3D reconstruction pipeline with the assumption that the camera and lens parameters do not change throughout the future data collection efforts.

The calibrated sensor system can be used to collect stereo video streams from a target scene. Data processing starts with extracting key video frames and is followed by detecting and matching feature points in different views. In order to find 3D coordinates of point $j$ in $k$-th

stereo view ( $p_{jk}$ ), the correct set of camera calibration parameters need to be determined first. For this purpose, the average of the following parameters are found from $\{\mathrm{P}_i \mid i = 1,...,n\}$: focal length ( $f_{avg}$ ), principal point ( $C_{avg}$ ), rotation ( $R_{avg}$ ), and translation ( $t_{avg}$ ). These average values are used to find an initial estimation for $Z$ coordinate of the point in the camera coordinate system (solve Eqs. 2-5 where $(u_1, v_1)$ and $(u_2, v_2)$ are the image coordinates of the point in the left and right views, respectively). To simplify and speed-up the calculations, the lens distortion effect is ignored here because only a very rough estimate of the $Z$ coordinate is needed in this step.

| | |
|---|---|
| $f_{(avg)x}\tilde{X} + (C_{(avg)x} - u_1)\tilde{Z} = 0$ | (2) |
| $f_{(avg)y}\tilde{Y} + (C_{(avg)y} - v_1)\tilde{Z} = 0$ | (3) |
| $(f_{(avg)x}R^{11}_{avg} + (C_{(avg)x} - u_2)R^{31}_{avg})\tilde{X} + (f_{(avg)x}R^{12}_{avg} + (C_{(avg)x} - u_2)R^{32}_{avg})\tilde{Y} +$ $(f_{(avg)x}R^{13}_{avg} + (C_{(avg)x} - u_2)R^{33}_{avg})\tilde{Z} = (u_2 - C_{(avg)x})t^{31}_{avg} - f_{(avg)x}t^{11}_{avg}$ | (4) |
| $(f_{(avg)y}R^{21}_{avg} + (C_{(avg)y} - v_2)R^{31}_{avg})\tilde{X} + (f_{(avg)y}R^{22}_{avg} + (C_{(avg)y} - v_2)R^{32}_{avg})\tilde{Y} +$ $(f_{(avg)y}R^{23}_{avg} + (C_{(avg)y} - v_2)R^{33}_{avg})\tilde{Z} = (v_2 - C_{(avg)y})t^{31}_{avg} - f_{(avg)y}t^{21}_{avg}$ | (5) |

The $\mathrm{P}_i$ that has the closest $D$ to the rough estimation of the $Z$ coordinate (i.e., $\tilde{Z}$ ) is selected as the correct camera calibration information for $p_{jk}$ . As can be inferred, the selected calibration parameters for two different points in a stereo view are not necessarily the same and they can change based on how far a point is from the camera set.

Once an appropriate $\mathrm{P}_i$ is selected for every point in a stereo pair, 3D coordinates of the points are calculated in the camera coordinate system by using a visual triangulation method.

Lens distortion parameters need to be included here to acquire more accurate estimations. These coordinates are then transformed into world coordinate system according to

$$\begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} = R_{cw}^T \begin{pmatrix} X_c - t_{cw} \\ Y_c - t_{cw} \\ Z_c - t_{cw} \end{pmatrix} \quad (6)$$

where $(X_w, Y_w, Z_w)^T$ is a point in the world coordinate system; $R_{cw}$ and $t_{cw}$ are the rotation matrix and translation vector from camera to world coordinate system; $(X_c, Y_c, Z_c)^T$ is the point in the camera coordinate system; and superscript $T$ represent the transpose of a matrix or vector.

The camera motion information and world coordinates of 3D points are finally refined in a bundle adjustment problem. This is a global non-linear optimization problem that minimizes a predefined cost function. The cost function $f(a,b) = \hat{x}$ is the Euclidean distance between the reprojection of a 3D point into 2D image space and 2D coordinates of the corresponding feature point that is detected in a video frame. The function $f$ takes $a = (a_1^T, a_2^T, ..., a_i^T)^T$ and $b = (b_1^T, b_2^T, ..., b_j^T)^T$ as input parameters and returns $\hat{x} = (\hat{x}_{p11}^T, ..., \hat{x}_{pij}^T)^T$. In this formulation, $a_i$ is the vector of estimated location for the left camera at time $i$; $b_j$ is the vector representing the 3D world coordinates of the $j$-th point; and $\hat{x}_{pij}$ is the projected image coordinates of world point $j$ in the $i$-th stereo frame.

An incremental 3D reconstruction approach is used to generate a 3D point cloud representing the entire scene. The reconstruction pipeline starts with the first stereo frame and the abovementioned process (i.e., initial estimation of the $Z$ coordinate of a point in a stereo frame, selection of an appropriate camera calibration set, calculation of the 3D point in camera

coordinate system, transformation of the 3D point into world coordinate system, and bundle adjustment optimization) is repeated once a new stereo frame is added.

## 5. Design of Experiments

This section provides details of experiments that are designed to validate the proposed stereo camera calibration procedure. The primary control variable in these experiments is the technical properties of the sensor system which need to be fixed while collecting the necessary data. A set of two video cameras capable of streaming raw video data and a pair of fixed focal length lenses are used. This is required to avoid the change of focal length and information loss during image compression. An appropriate baseline distance is also selected based on the typical range values that are encountered in infrastructure applications (typical range values are 10-25m). Once the sensor system is ready, the following parameters should not change while collecting the necessary data: video resolution, focal length, and relative position of the two cameras. The baseline distance between the left and right cameras ($b$) can be selected based on a simple formulation presented in (Gallup, 2011) for analyzing the reconstruction accuracy in a stereo setup.

$$\varepsilon_z = \frac{bf}{d} - \frac{bf}{d+\varepsilon_d} = \frac{z^2\varepsilon_z}{bf + z\varepsilon_d} \approx \frac{z^2}{bf} \times \varepsilon_d \tag{7}$$

where $z$ is the depth in cm, $\varepsilon_z$ is the expected measurement error in cm, $f$ is the focal length measured in pixels, and $\varepsilon_d$ is the disparity error of a feature correspondence. As an example, in case of using two 5MP cameras with 16mm fixed focal length lenses and assuming $z = 20m$, $\varepsilon_z = 2cm$, $f = 6500$, and $\varepsilon_d = 0.1$, the baseline distance can be calculated as 30cm.

A checkerboard with an appropriate number of black and white squares in two perpendicular directions is also required for the calibration process. The number of squares and their dimensions are selected according to the scene (a pattern of 13×14 squares each with a dimension of 60mm).

Two sets of experiments were designed to study the impact of the conventional and proposed calibration procedures on the accuracy of 3D coordinates of points. The first experiment includes 3D reconstruction of a building façade with intersecting planar faces. The planes are well-textured and have a brick pattern. The second experiment includes another building façade with planar faces, but covered with poorly-textured aluminum panels. Fig. 3 shows a snapshot of the two environments. The scenes are selected to be planar for a main reason: the planarity allows controlling the $Z$ coordinate of 3D points in the desired range by simply changing the distance between the stereo camera system and the planar face.

For camera calibration, six sets of stereo video streams are captured from the board under different conditions. In the first set which will be used for testing conventional procedures, the

**Figure 3:** Two building façades for multi-step stereo camera calibration experiments

depth of calibration changes in the range of $5m \le D_1 \le 15m$ while capturing the videos. Captured video frames should cover different views and angles of the board while the camera moves smoothly toward and/or away from the board. The next five sets are needed to test the proposed stereo camera calibration procedure. In these sets, the distance between the depth of calibration is fixed to $D_2 = 5m$, $D_3 = 10m$, $D_4 = 15m$, $D_5 = 20m$, and $D_6 = 25m$, respectively. These limits have been selected according to the typical range values that we encounter in building applications. The sensor system is also used to collect stereo videos from the planar scenes while the distance of the camera to the planar scenes changes from $5m \le D \le 25m$. This data is a control variable and will be used for 3D reconstruction of the scenes in two scenarios: a) using conventional calibration procedures (parameters acquired from the 1st set of calibration videos); and b) using the proposed multi-step calibration procedure (multiple sets of parameters acquired from the 2nd to 6th set of calibration videos).

The performance of the proposed calibration procedure is assessed based on the following metrics: a) spatial distance accuracy of the initial estimation for 3D coordinates of points with different range values (only one set of stereo frames is used in this case); and b) spatial distance accuracy of a dense 3D point cloud. For the first metric, stereo frames corresponding to $D = \{5,10,15,20,25m\}$ are extracted from the videos to detect and match feature points. Calibration parameters acquired from the conventional and proposed procedure are then used to estimate 3D coordinates of feature points from left and right views of stereo frames. Spatial distance between pairs of feature points is then calculated for each case and compared to the ground truth data that is acquired using total station surveying. For the second metric, calibration parameters from the conventional and proposed procedure are used separately in a dense 3D reconstruction package and the spatial accuracy of the results is evaluated. The sample

size at all experiments is considered to be 384 which correspond to 95% confidence level and ±5% confidence interval.

## 6. Implementation and Results

A prototype was created using Microsoft Visual C# to implement and test the proposed multi-step calibration procedure and subsequent 3D reconstruction steps. The C# platform provides a base to connect to any number of cameras with real-time responsiveness. OpenCV (Intel® Open Source and free C++ Computer Vision Library) was selected as the main image processing library. Two high resolution Flea2 cameras were used to capture stereo video streams. The baseline distance was approximately 30cm and the video resolution was 5MP with a frame rate of 7.5 fps. A calibration board with a pattern of 13×14 squares each with a dimension of 60mm was also built.

An automatic stereo camera calibration software was developed using the functions available in OpenCV. The user runs the program while videotaping a calibration pattern at a predefined distance from the camera set. The program is real-time responsive and automatically detects the calibration pattern in every video frame. Once the pattern is successfully detected in a stereo frame using the OpenCV's *cvFindChessboardCorners* function, chessboard corners are automatically refined to their location with subpixel accuracy and also matched between the two views by invoking the *cvFindCornerSubPix* function (Fig. 4). This process continues until enough number of views are captured (typically between 30 to 40). Then, the calibration function (*cvStereoCalibrate*) is invoked and the necessary parameters are calculated. *cvStereoCalibrate* provides the possibility of calibrating a stereo camera set according to different constraints such as zero radial or tangential distortions, fixed principal point, fixed aspect ratio, and/or fixed focal length. The same process is repeated for different *D* values.

**Figure 4:** Automatically detected and matched calibration board corners

The designed experiments were performed according to the specified details. The previously mentioned camera system and calibration board were used to capture the six sets of required data for calibration. Using the developed automatic calibration software, 50 stereo frames were extracted in each case (i.e., $D = 5m$, $D = 10m$, $D = 15m$, $D = 20m$, and $D = 25m$) and the calibration parameters were calculated. Fig. 5 demonstrates some of the intermediate results. Then, two sets of stereo video streams were captured from the façade with brick pattern and the façade with aluminum panels while the distance between the camera system and planar faces was changing in the range of $5m \le D \le 25m$.

For evaluating the first performance metric (i.e., spatial accuracy of the initial estimation for 3D coordinates of points with different range values), stereo frames corresponding to $D = \{5,10,15,20,25m\}$ were extracted from the façade videos and 3D coordinates of feature points were calculated using the sets of estimated calibration parameters. Spatial distance between pairs of 3D feature points was then compared to the ground truth data. Table 2 illustrates the average error at each scenario (sample size of 384). The results indicate that a more accurate initial estimation can be acquired for a point at a range of $Z$ using the calibration parameters that correspond to $D \approx Z$; this supports the hypothesis presented in this paper.

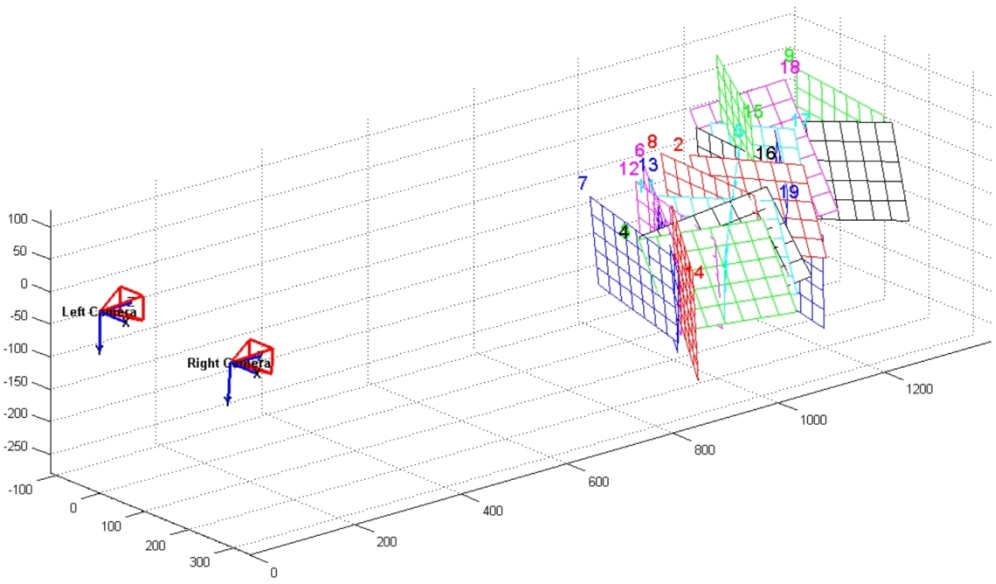



**Figure 5:** Visualization of the extrinsic parameters in stereo camera calibration

**Table 2:** Average spatial distance error for different calibration scenarios

| **Calibration Scenario** | **Average spatial distance error (cm)** | | | | |
|---|---|---|---|---|---|
| | $Z \approx 5m$ | $Z \approx 10m$ | $Z \approx 15m$ | $Z \approx 20m$ | $Z \approx 25m$ |
| $D = 5m$ | **±2.5** | ±4.7 | ±9.9 | ±17.5 | ±23.7 |
| $D = 10m$ | ±2.8 | **±4.4** | ±8.4 | ±15.5 | ±20.3 |
| $D = 15m$ | ±3.6 | ±5.2 | **±6.4** | ±14.6 | ±19.8 |
| $D = 20m$ | ±4.4 | ±6.0 | ±9.1 | **±11.3** | ±18.3 |
| $D = 25m$ | ±5.0 | ±7.6 | ±12.5 | ±15.1 | **±15.2** |
| $5m \leq D \leq 25m$ | ±3.3 | ±4.9 | ±10.3 | ±15.8 | ±19.0 |

To evaluate the second performance metric (i.e., spatial distance accuracy of a dense 3D point cloud), two dense 3D point clouds were generated at each experiments: one using the information acquired from a conventional stereo camera calibration algorithm and another using the proposed procedure. The key-frame selection method proposed in (Rashidi, et al., 2013) was used to extract frames that have minimum motion blur and appropriate number of feature points while the camera motion between two consecutive key-frames is larger than a minimum

threshold. In addition, a modified version of the patch-based multi-view stereo software, which is based on (Furukawa & Ponce, 2010) and available online, was used to generate the dense 3D reconstructions. 384 pairs of points were selected randomly from the generated dense 3D point clouds and their spatial distance was compared to the ground truth data. Total station surveying was used to acquire the ground truth data.

Fig. 6 demonstrates the results involving the building façade with a brick pattern (i.e., the first experiment). The 95 percentile error in the point cloud generated from the information acquired by conventional calibration algorithms was ±12.8cm while this error was ±9.6cm in the point cloud generated by using the proposed calibration procedure. This shows a reduction of 3.2cm (25%) in the spatial distance error because of using the proposed multi-step stereo camera calibration procedure. The relative improved accuracy can also be visually seen by comparing the point clouds in Fig. 6(a) and Fig. 6(b). The second point cloud is sharper in planar areas. This supports the presented hypothesis in this paper. It is necessary to mention that this accuracy may be further improved by modifying the multi-view geometry process which is out of the scope of this paper.

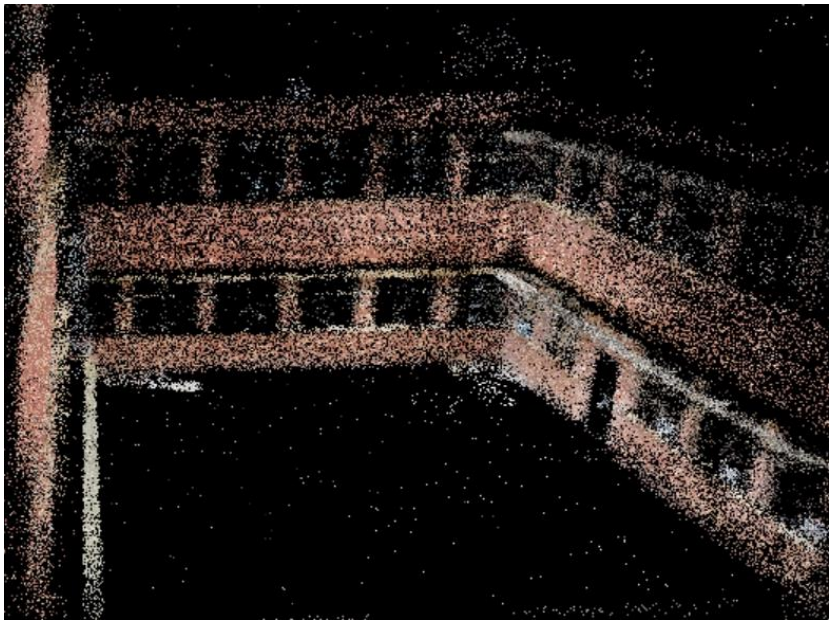
Fig. 7 shows the generated dense 3D point clouds for the second experiment (i.e.,

building façade with aluminum panels). Fig. 7(a) demonstrates the 3D point cloud acquired by

using the conventional stereo camera calibration algorithms. Fig. 7(b) shows the same point

cloud when the proposed multi-step calibration procedure is used. No significant difference in

the appearance of the two point clouds can be noted in the front view snapshots. However, the



**(a)**

**(b)**

**Figure 6:** Dense 3D point cloud of the building façade with a brick pattern. (a) Conventional calibration method. (b) Multi-step calibration procedure.

quality of these point clouds can be visually evaluated when the variance of the points on the

planar surface is examined. This variance is demonstrated in Fig. 7(c) and Fig. 7(d). As can be seen, in case of using the multi-step calibration procedure, the variance of the points on the planar surface is much lower than the case that uses conventional calibration algorithms. In general, this is a very challenging scene to be reconstructed with image-based 3D reconstruction algorithms due to the prevalence of repetitive patterns and also existence of poorly-textured

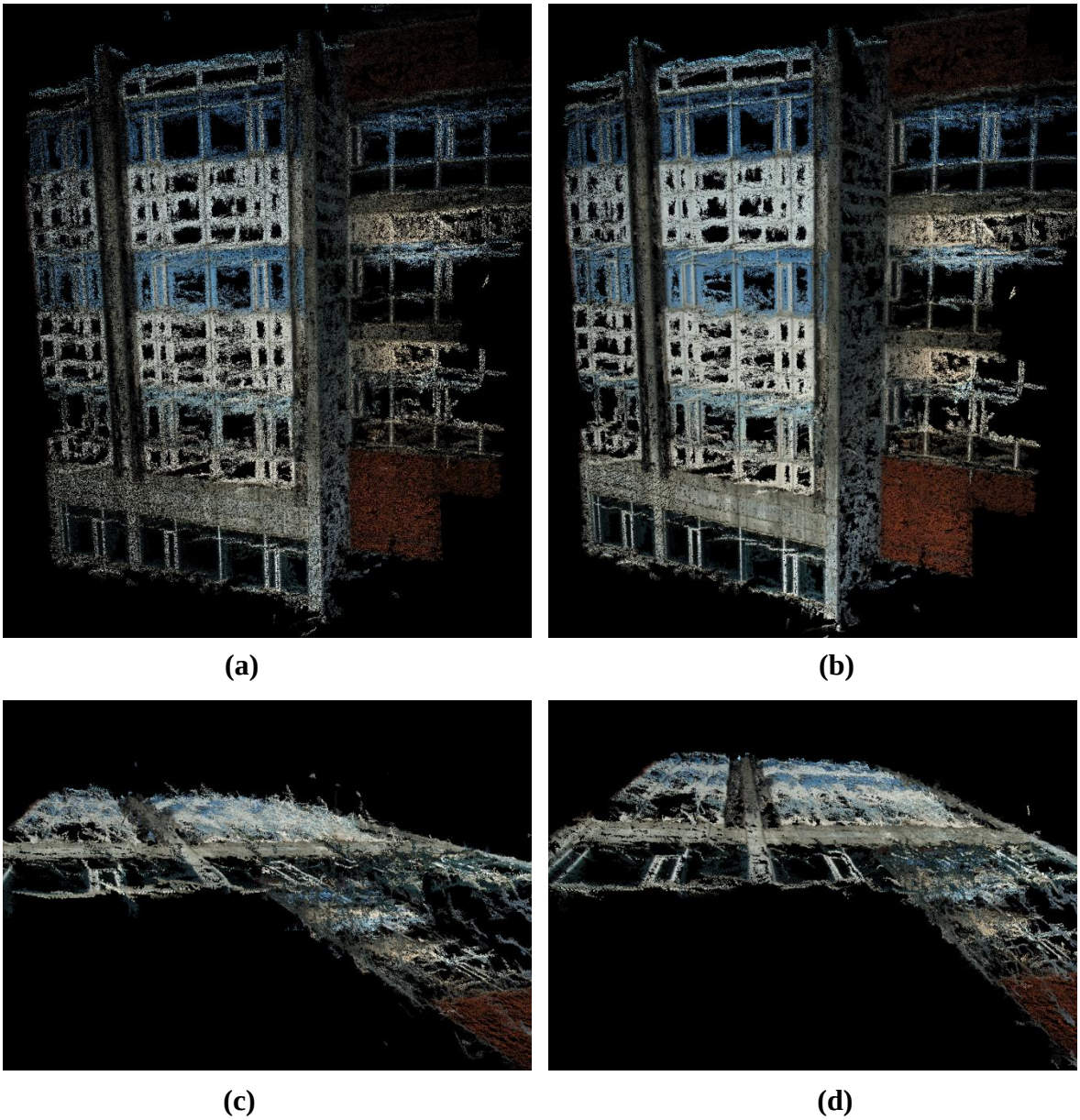



**Figure 7:** Dense 3D point cloud of the building façade with aluminium panels. (a,c) Conventional calibration method. (b,d) Multi-step calibration procedure

surfaces. The Euclidean accuracy of the coordinates of 3D points were therefore less than the ones acquired in the previous experiment. The 95 percentile spatial distance errors in this case were ±16.5cm for conventional methods and ±11.7cm for the proposed procedure. This indicates an average reduction of 4.8cm (29%) in spatial distance errors.

## 7. Conclusion Remarks

Accurate 3D reconstruction of infrastructure from multiple-view imagery can provide the construction industry with an inexpensive alternative to the laser-based surveying techniques. In the case of using a calibrated stereo camera system, several observations have shown that the accuracy of final results is very sensitive to the calibration parameters especially in far-range applications. The highest sensitivity corresponds to the distortion coefficients. Due to this sensitivity, the existing stereo camera calibration algorithms only provide accurate results when they are used in close-range applications.

This paper presented a novel multi-step stereo camera calibration procedure to alleviate the abovementioned problem. The goal was to enhance the Euclidean accuracy of the generated dense 3D point clouds in far-range scenarios. The proposed procedure uses a set of discrete values to represent the distance between the sensor system and the calibration board ($D$). For each $D$, a set of stereo video streams are collected while the distance between the camera and the board is fixed to $D$. Conventional stereo camera calibration algorithms are then used to calculate calibration parameters for the given $D$. Repeating this process for all the values results in multiple sets of camera parameters each corresponding to a specific $D$. These sets are then used in the SfM process with the following assumption: for each 3D point, the set of calibration parameters that have the closest $D$ value to the point's $Z$ coordinate are used. Results from two different case studies demonstrated that this procedure is capable of reducing the spatial

measurement errors by 25% in 3D reconstruction of a building façade with a brick pattern and 29% in 3D reconstruction of a building façade with aluminum panels. As mentioned before, camera calibration and dense multi-view geometry are key issues regarding the capability to achieve spatial accuracy levels that could compete with laser-based spatial data collection systems.

## 8. References


Dai, F., Rashidi, A., Brilakis, I. & Vela, P., 2013. Comparison of image- and time-of-flight-based technologies for 3D reconstruction of infrastructure. *Journal of Construction Engineering and Management,* 139(1), pp. 69-79.

Dang, T., Hoffmann, C. & Stiller, C., 2009. Continuous stereo self-calibration by camera parameter tracking. *IEEE Transactions on Image Processing,* 18(7), pp. 1536-1550.

Fathi, H. & Brilakis, I., 2011. Automated sparse 3D point cloud generation of infraastructure using its distinctive visual features. *Advanced Engineering Informatics,* 25(4), pp. 760-770.

Fathi, H. & Brilakis, I., 2013. A videogrammetric as-built data collection method for digital fabrication of sheet metal roof panels. *Advanced Engineering Informatics,* 27(4), pp. 466-476.

Furukawa, Y. & Ponce, J., 2009. Accurate camera calibration from multi-view stereo and bundle adjustment. *International Journal of Computer Vision,* 84(3), pp. 257-268.

Furukawa, Y. & Ponce, J., 2010. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 32(8), p. 1362–1376.

Gallup, D., 2011. *Efficient 3D reconstruction of large-scale urban environments from street-level video,* s.l.: Dissertation, University of North Carolina.

Geiger, A., Zeigler, J. & Stiller, C., 2011. *StereoScan: dense 3D reconstruction in real-time.* s.l., IEEE Intelligent Vehicles Symposium.

Golparvar-Fard, M., Peña-Mora, F. & Savarese, S., 2014. Automated progress monitoring using unordered daily construction photographs and IFC-based building information models. *Journal of Computing in Civil Engineering,* Volume in press, pp. doi:10.1061/(ASCE)CP.1943-5487.0000205.

Hartley, R. & Kang, S., 2005. *Parameter-free radial distortion correction with centre of distortion estimation.* s.l., International Conference on Computer Vision.

Hartley, R. & Zisserman, A., 2003. *Multiple view geometry in computer vision.* 2nd ed. Cambridge: Cambridge University Press.

He, B. & Li, Y., 2008. Camera calibration from vanishing points in a vision system. *Optics and Laser Technology,* Volume 40, pp. 555-561.

House, B. & Nickels, K., 2006. Increased automation in stereo camera calibration techniques. *Systemics, Cybernetics and Informatics,* 4(4), pp. 48-51.

Kim, J., Gurdjos, P. & Kweon, I., 2005. Geometric and algebraic constraints of projected concentric circles and their applications to camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27(4), pp. 637-642.

Kim, J. & Kweon, I., 2009. Camera calibration based on arbitrary parallelograms. *Computer Vision and ImageUnderstanding,* Volume 113, pp. 1-10.

Klein, L., Li, N. & Becerik-Gerber, B., 2012. Image-based verification of as-built documentation of operational buildings. *Automation in Construction,* Volume 21, pp. 161-171.

Liebowitz, D., 2001. *Camera calibration and reconstruction of geometry from images,* s.l.: PhD Thesis, University of Oxford.

Mavrinac, A., Chen, X. & Tepe, K., 2010. An automatic calibration method for stereo-based 3D distributed smart camera networks. *Computer Vision and Image Understanding,* Volume 114, pp. 952-962.

Peng, J., 2011. *Comparison of three dimensional measurement accuracy using stereo vision,* s.l.: Thesis, University of Regina.

Pollefeys, M. et al., 2008. Detailed real-time urban 3D reconstruction from video. *International Journal of Computer Vision,* 78(2-3), pp. 143-167.

Rashidi, A., Dai, F., Brilakis, I. & Vela, P., 2013. Optimized selection of key frames for monocular videogrammetric surveying of civil infrastructure. *Advanced Engineering Informatics,* 27(2), pp. 270-282.

Ricolfe-Viala, C. & Sanchez-Salmeron, A., 2010. Robust metric calibration of non-linear camera lens distortion. *Pattern Recognition,* Volume 43, pp. 1688-1699.

Seitz, S. et al., 2006. *A comparison and evaluation of multi-view stereo reconstruction algorithms.* s.l., IEEE Conference on CVPR.

Strecha, C. et al., 2008. *On benchmarking camera calibration and multi-view stereo for high resolution imagery.* s.l., IEEE Conference on Computer Vision and Pattern Recognition.

Strecha, C. et al., 2008. *On benchmarking camera calibration and multi-view stereo for high resolution imagery.* s.l., IEEE Conference on Computer Vision and Pattern Recognition.

Sturm, P. & Ramalingam, S., 2004. *A generic concept for camera calibration.* s.l., Fifth European Conference on Computer Vision.

Wang, J., Shi, F., Zhang, J. & Liu, Y., 2008. A new calibration model of camera lens distortion. *Pattern Recognition,* 41(2), pp. 607-615.

Wu, C., 2013. *Towards linear-time incremental structure from motion.* Seattle, IEEE International Conference on 3D Vision.

Xiao, Z., Jin, L., Yu, D. & Tang, Z., 2010. A cross-target-based accurate calibration method of binocular stereo systems with large-scale field of view. *Measurement,* Volume 43, pp. 747-754.

Xu, G., Chen, L. & Gao, F., 2012. *Study on binocular stereo camera calibration method.* s.l., International Conference on IASP.

Zhang, Z., 2000. A flexible new technique for camera calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* 22(11), pp. 1330-1334.

Zhang, Z., 2004. Camera calibration with one-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26(7), pp. 892-899.