Accurate and Practical Calibration of a Depth and Color Camera Pair

Daniel Herrera C., Juho Kannala, and Janne Heikkilä

Machine Vision Group
University of Oulu
{dherrera,jkannala,jth}@ee.oulu.fi

Abstract. We present an algorithm that simultaneously calibrates a color camera, a depth camera, and the relative pose between them. The method is designed to have three key features that no other available algorithm currently has: accurate, practical, applicable to a wide range of sensors. The method requires only a planar surface to be imaged from various poses. The calibration does not use color or depth discontinuities in the depth image which makes it flexible and robust to noise. We perform experiments with particular depth sensor and achieve the same accuracy as the propietary calibration procedure of the manufacturer.

Keywords: calibration, depth camera, camera pair

1 Introduction

Obtaining depth and color information simultaneously from a scene is both highly desirable and challenging. Depth and color are needed in applications ranging from scene reconstruction to image based rendering. Capturing both simultaneously requires using two or more sensors. A basic device for scene reconstruction is a depth and color camera pair. Such a camera pair consists of a color camera rigidly attached to a depth sensor (e.g. time-of-flight (ToF) camera, laser range scanner, structured light scanner).

In order to reconstruct a scene from the camera pair measurements the system must be calibrated. This includes internal calibration of each camera as well as relative pose calibration between the cameras. Color camera calibration has been studied extensively [1, 2] and different calibration methods have been developed for different depth sensors. However, independent calibration of the cameras may not yield the optimal system parameters, and a comprehensive calibration of the system as a whole could improve individual camera calibration as it allows to use all the available information.

Previous work. A standard approach is to calibrate the cameras independently and then calibrate only the relative pose between them [3–5]. This may not be the optimal solution as measurements from one camera can improve the calibration of the other camera. Moreover, the independent calibration of a depth camera can require a high precision 3D calibration object that can be avoided using joint calibration.

Fuchs and Hirzinger [6] propose a multi-spline model for ToF cameras. Their model has a very high number of parameters and it requires a robotic arm to

know the exact pose of the camera. Lichti [7] proposes a calibration method for an individual laser range scanner using only a planar calibration object. It performs a comprehensive calibration of all parameters. However, it relies on the varying response of the scanner to different surface colors to locate corner features on the image.

Zhu et al. [8] describe a method for fusing depth from stereo cameras and ToF cameras. Their calibration uses the triangulation from the stereo cameras as ground truth. This ignores the possible errors in stereo triangulation and measurement uncertainties. The different cameras are thus calibrated independently and the parameters obtained may not be optimal.

Motivation. As a motivation for our work, we propose three requirements that an optimal calibration algorithm must have. To the best of our knowledge, no available calibration algorithm for a depth and color camera pair fulfills all three criteria.

Accurate: The method should provide the best combination of intrinsic and extrinsic parameters that minimizes the reprojection error for both cameras over all calibration images. This may seem like an obvious principle but we stress it because partial calibrations, where each camera is calibrated independently and the relative pose is estimated separately, may not achieve the best reprojection error.

Practical: The method should be practical to use with readily available materials. A high precision 3D calibration object is not easy/cheap to obtain and a robotic arm or a high precision mechanical setup to record the exact pose of the camera pair is usually not practical, whereas a planar surface is usually readily available.

Widely applicable: To be applicable to a wide range of depth sensors, one cannot assume that color discontinuities are visible on the depth image. Moreover, some depth sensors, like the one used for our experiments, may not provide accurate measurements at sharp depth discontinuities. Thus, neither color nor depth discontinuities are suitable features for depth camera calibration. The method should use features based on depth measurements that are most reliable for a wide range of cameras.

2 The depth and color camera pair

Our setup consists of one color camera and one depth sensor rigidly attached to each other. Our implementation and experiments use the Kinect sensor from Microsoft, which consists of a projector-camera pair as the depth sensor that measures per pixel disparity. The Kinect sensor has gained much popularity in the scientific and the entertainment community lately. The complete model includes $20 + 6N$ parameters where N is the number of calibration images. The details of the model are described below.

Color camera intrinsics. We use a similar intrinsic model as Heikkilä and Silven [1] which consists of a pinhole model with radial and tangential distortion correction. The projection of a point from color camera coordinates $\mathbf{x}_c = [x_c, y_c, z_c]^\top$ to color image coordinates $\mathbf{p}_c = [u_c, v_c]^\top$ is obtained through

the following equations. The point is first normalized by $\mathbf{x}_n = [x_n, y_n]^\top = [x_c/z_c, y_c/z_c]^\top$. Distortion is then performed:

$$\mathbf{x}_g = \begin{bmatrix} 2k_3 x_n y_n + k_4(r^2 + 2x_n^2) \\ k_3(r^2 + 2y_n^2) + 2k_4 x_n y_n \end{bmatrix} \tag{1}$$

$$\mathbf{x}_k = (1 + k_1 r^2 + k_2 r^4)\mathbf{x}_n + \mathbf{x}_g \tag{2}$$

where $r^2 = x_n^2 + y_n^2$ and k is a vector containing the four distortion coefficients.

Finally the image coordinates are obtained:

$$\begin{bmatrix} u_c \\ v_c \end{bmatrix} = \begin{bmatrix} f_{cx} & 0 \\ 0 & f_{cy} \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} u_{c0} \\ v_{c0} \end{bmatrix} \tag{3}$$

The complete color model is described by $\mathcal{L}_c = \{f_{cx}, f_{cy}, u_{c0}, v_{c0}, k_1, k_2, k_3, k_4\}$.

Depth camera intrinsics. In our experiments we used the increasingly popular Kinect sensor as a depth camera [9]. However, the method allows any kind of depth sensor to be used by replacing this intrinsic model. The Kinect consists of an infrared projector that produces a constant pattern and a camera that measures the disparity between the observed pattern and a pre-recorded image at a known constant depth. The output consists of an image of scaled disparity values.

The transformation between depth camera coordinates $\mathbf{x}_d = [x_d, y_d, z_d]^\top$ and depth image coordinate $\mathbf{p}_d = [u_d, v_d]$ follows the same model used for the color camera. The distortion correction did not improve the reprojection error and the distortion coefficients were estimated with very high uncertainty. Therefore we do not use distortion correction for the depth image.

The relation between the disparity value d and the depth z_d is modeled using the equation:

$$z_d = \frac{1}{\alpha(d - \beta)} \tag{4}$$

where α and β are part of the depth camera intrinsic parameters to be calibrated. The model for the depth camera is described by $\mathcal{L}_d = \{f_{dx}, f_{dy}, u_{d0}, v_{d0}, \alpha, \beta\}$.

Extrinsics and relative pose. Figure 1 shows the different reference frames present in a scene. Points from one reference frame can be transformed to another using a rigid transformation denoted by $\mathcal{T} = \{\mathbf{R}, \mathbf{t}\}$, where $\mathbf{R}$ is a rotation and $\mathbf{t}$ a translation. For example, the transformation of a point $\mathbf{x}_w$ from world coordinates $\{W\}$ to color camera coordinates $\{C\}$ follows $\mathbf{x}_c = \mathbf{R}_c\mathbf{x}_w + \mathbf{t}_c$. Reference $\{V\}$ is anchored to the corner of the calibration plane and is only used for initialization. The relative pose $\mathcal{T}_r$ is constant, while each image has its own pose $\mathcal{T}_c$, resulting in $6 + 6N$ pose parameters.

3 Calibration method

We use a planar checkerboard pattern for calibration which can be constructed from any readily available planar surface (e.g. a flat table, a wall). The checkerboard corners provide suitable constraints for the color images, while the planarity of the points provides constraints on the depth image. The pixels at the

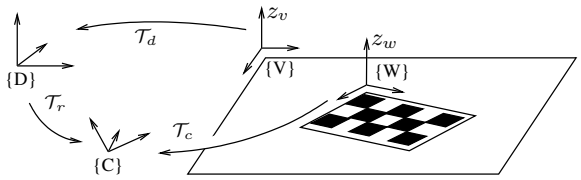


Fig. 1: Reference frames and transformations present on a scene. $\{C\}$ and $\{D\}$ are the color and depth cameras' reference frames respectively. $\{V\}$ is the reference frame anchored to the calibration plane and $\{W\}$ is the world reference frame anchored to the calibration pattern.

borders of the calibration object can be ignored and thus depth discontinuities are not needed. Figure 2 shows a sample image pair used for calibration. Figure

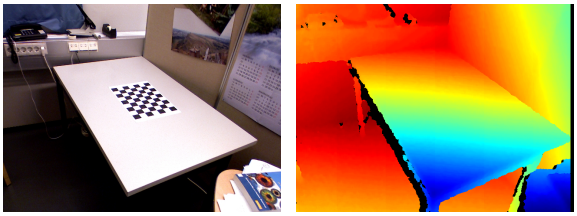


Fig. 2: Sample calibration images. Note the inaccuracies at the table's edge.

3 shows the steps of the calibration and its inputs. An initial estimation for the calibration parameters is obtained by independently calibrating each camera. The depth intrinsic parameters $\mathcal{L}_d$ and the relative pose $\mathcal{T}_r$ are then refined using a non-linear optimization. Finally, all parameters are refined simultaneously.

Corner based calibration. The calibration of a color camera is a well studied problem, we use Zhang's method [2, 10] to initialize the camera parameters. Briefly, the steps are the following. The checkerboard corners are extracted from the intensity image. A homography is then computed for each image using the known corner positions in world coordinates $\{W\}$ and the measured positions in the image. Each homography then imposes constraints on the camera parameters which are then solved with a linear system of equations. The distortion coefficients are initially set to zero.

The same method is used to initialize the depth camera parameters. However, because the checkerboard is not visible in the depth image, the user selects the four corners of the calibration plane (the whole table in figure 2). These corners are very noisy and are only used here to obtain an initial guess. The homography is thus computed between $\{V\}$ and $\{D\}$. This initializes the focal lengths,

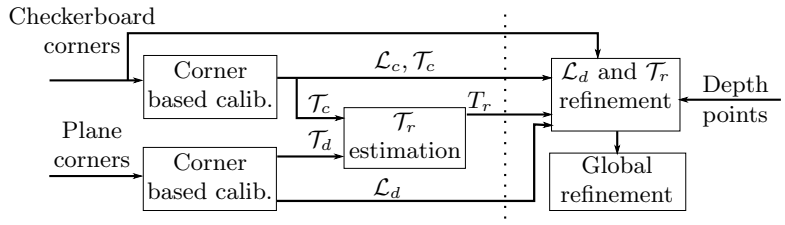


Fig. 3: Block diagram of the calibration algorithm. **Left of dashed line:** initialization. **Right of dashed line:** non-linear minimization.

principal point, and the transformation $\mathcal{T}_d$. Using these initial parameters we obtain a guess for the depth of each selected corner. With this depth and the inverse of the measured disparity an overdetermined system of linear equations is built using (4), which gives an initial guess for the depth parameters (α and β).

Relative pose estimation. The independent calibrations give an estimation of the transformations $\mathcal{T}_c$ and $\mathcal{T}_d$. However, the reference frames $\{W\}$ and $\{V\}$ are not aligned. By design we know that they are coplanar. We can use this information by extracting the plane equation in each reference frame and using it as a constraint. We define a plane using the equation $\mathbf{n}^\top \mathbf{x} - \delta = 0$ where $\mathbf{n}$ is the unit normal and δ is the distance to the origin.

If we divide a rotation matrix into its colums $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$ and know that the parameters of the plane in both frames are $\mathbf{n} = [0, 0, 1]^\top$ and $\delta = 0$, the plane parameters in camera coordinates are:

$$\mathbf{n} = \mathbf{r}_3 \quad \text{and} \quad \delta = \mathbf{r}_3^\top \mathbf{t} \tag{5}$$

where we use $\mathbf{R}_c$ and $\mathbf{t}_c$ for the color camera and $\mathbf{R}_d$ and $\mathbf{t}_d$ for the depth camera.

As mentioned by Unnikrishnan and Hebert [4] the relative pose can be obtained in closed form from several images. The plane parameters for each image are concatenated in matrices of the form: $\mathbf{M}_c = [\mathbf{n}_{c1}, \mathbf{n}_{c2}, ..., \mathbf{n}_{cn}]$, $\mathbf{b}_c = [\delta_{c1}, \delta_{c2}, ..., \delta_{cn}]$, and likewise for the depth camera to form $\mathbf{M}_d$ and $\mathbf{b}_d$. The relative transformation is then:

$$\mathbf{R}'_r = \mathbf{M}_d \mathbf{M}_c^\top \quad \text{and} \quad \mathbf{t}_r = (\mathbf{M}_c \mathbf{M}_c^\top)^{-1} \mathbf{M}_c (\mathbf{b}_c - \mathbf{b}_d)^\top \tag{6}$$

Due to noise $\mathbf{R}'_r$ may not be orthonormal. We obtain a valid rotation matrix through SVD using: $\mathbf{R}_r = UV^\top$ where $USV^\top$ is the SVD of $\mathbf{R}'_r$.

Non-linear minimization. The calibration method aims to minimize the weighted sum of squares of the measurement reprojection errors. The error for the color camera is the Euclidean distance between the measured corner position $\mathbf{p}_c$ and its reprojected position $\mathbf{p}'_c$. Whereas for the depth camera it is the difference between the measured disparity d and the predicted disparity d' obtained by inverting (4). Because the errors have different units, they are weighted using

the inverse of the corresponding measurement variance, σ_c^2 and σ_d^2. The resulting cost function is:

$$c = \sigma_c^{-2} \sum \left[(u_c - u'_c)^2 + (v_c - v'_c)^2 \right] + \sigma_d^{-2} \sum (d - d') \tag{7}$$

Note that (7) is highly non-linear. The Levenberg-Marquardt algorithm is used to minimize (7) with respect to the calibration parameters. The initialization gives a very rough guess of the depth camera parameters and relative pose, whereas the color camera parameters have fairly good initial values. To account for this, the non-linear minimization is split in two phases. The first phase uses fixed parameters for the color camera $\mathcal{L}_c$ and external pose $\mathcal{T}_c$, and optimizes the depth camera parameters $\mathcal{L}_d$ and the relative pose $\mathcal{T}_r$. A second minimization is performed over all the parameters to obtain an optimal estimation.

Variance estimation. An initial estimate of the color measurement variance σ_c^2 is estimated from the residuals after the first independent calibration. An estimate of the disparity variance σ_d^2 is obtained from the disparity residuals after the first non-linear minimization. It is noted that, because $\mathcal{L}_c$ and $\mathcal{T}_c$ are fixed, the color residuals do not need to be computed and σ_d^2 plays no role in this minimization. The second minimization stage, when all parameters are refined, is then run iteratively using the previously obtained residual variances as the measurement variances for the next step until they converge.

4 Results

We tested our calibration method with an off-the-shelf Kinect device. The device consists of a color camera, an infrared camera and an infrared projector arranged horizontally. The electronics of the device compute a depth map for the infrared image based on the observed pattern from the projector. We ignore the infrared image and use only the depth information and treat it as a generic depth and color camera pair. We used a dataset of 55 images, 35 were used for calibration and 20 for validation. Both sets cover similar depth ranges (0.5m to 2m) and a wide range of poses. For the validation set, (7) was minimized only over the external pose $\mathcal{T}_c$ to find the best pose for the previously obtained calibration.

Parameters and residuals. The obtained calibration parameters and their uncertainties are presented in Table 1. Figure 4 presents histograms of the residuals for the validation set. The formulation of our cost function (7) allows us to use the uncertainty analysis presented by Hartley and Zisserman [11]. They show that the covariance of the estimated parameters Σ_P can be obtained directly from the Jacobian of the cost function J and the covariance of the measurements Σ_X using:

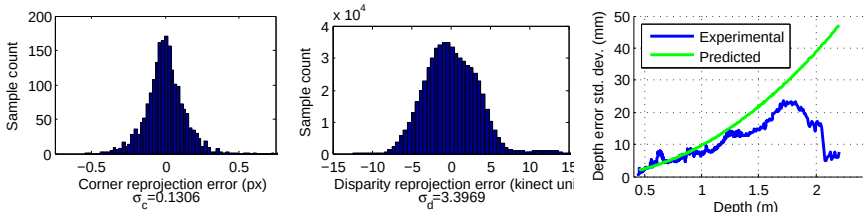
$$\Sigma_P = \left(J^\top \Sigma_X J \right)^{-1} \tag{8}$$

Depth uncertainty. The disparity errors are well modeled by a gaussian distribution. Using (4) and the estimated disparity variance, we obtained numerically the expected variance in depth for each disparity value. Separate statistics are computed for each depth present in the validation set to obtain an experimental

Table 1: Obtained calibration parameters. Error estimates correspond to three times their standard deviation.

Color internals							
f_{cx}	f_{cy}	u_{c0}	v_{c0}	k_1	k_2	k_3	k_4
532.90	531.39	318.57	262.08	0.2447	-0.5744	0.0029	0.0065
±0.06	±0.05	±0.07	±0.07	±0.0004	±0.0017	±0.0001	±0.0001

Depth internals						Relative pose (rad, mm)			
f_{dx}	f_{dy}	u_{d0}	v_{d0}	α	β	θ_r	t_{rx}	t_{ry}	t_{rz}
593.36	582.74	322.69	231.48	-0.00285	1091.0	0.024	-21.4	0.7	1.0
±1.81	±2.48	±1.34	±1.59	±0.00001	±1.0	±0.003	±1.5	±1.5	±1.9



(a) Color residuals (b) Depth residuals (c) Depth uncertainty

Fig. 4: Obtained error residuals and depth uncertainty.

depth variance. Both curves are shown in Figure 4c. The experimental curve shows the expected increase in variance as the depth increases. The final drop in variance is due to low sample count at the end of the range.

Comparison with manufacturer calibration. The manufacturer of the Kinect sensor, PrimeSense, has a proprietary camera model and calibration procedure. They provide an API to convert the disparity image to a point cloud in world coordinates. To validate our calibration against the one from the manufacturer, we took an image from a slanted planar surface that covers a range of depths. The disparity image was reprojected to world coordinates using our model and the manufacturer's API. A plane was fitted to each point cloud and the distance of the points to the plane was computed. The manufacturer's reprojection had a standard deviation of 3.10mm from the plane, while ours was 3.00mm. This proves that our calibration of the depth camera has comparable accuracy to that of the manufacturer.

Colorized point cloud. The fully calibrated system can be used to obtain a colored point cloud in metric coordinates. For illustration purposes, Figure 5 shows an example scene and a reprojection from a different view point.

5 Conclusions

The results show that our algorithm performed adequately for the chosen camera pair. In addition, we believe that our algorithm is flexible enough to be used with other types of depth sensors by replacing the intrinsics model of the depth

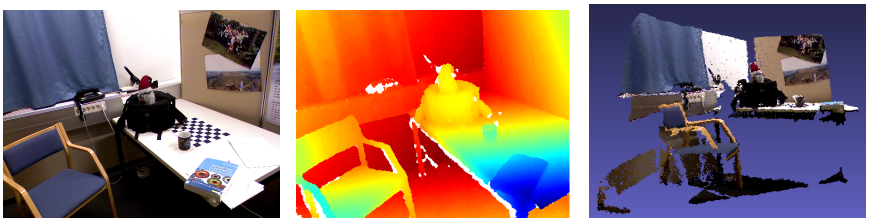


Fig. 5: Sample scene. Color image, depth map, and change of view point.

camera. The constraints used can be applied to any type of depth sensor. Future work can include the calibration of a ToF and color camera pair.

We have presented a calibration algorithm for a depth and color camera pair that is optimal in the sense of the postulated principles. The algorithm takes into account color and depth features simultaneously to improve calibration of the camera pair system as a whole. It requires only a planar surface and a simple checkerboard pattern. Moreover, the method is flexible to be used with different types of depth sensors. Finally, our method showed comparable accuracy to the one provided by the manufacturer of a particular depth camera.

Acknowledgements. This project has been funded by the Academy of Finland's project #127702.

References

1. J. Heikkilä and O. Silven, "A Four-step Camera Calibration Procedure with Implicit Image Correction," in *CVPR*. IEEE, 1997, p. 1106.
2. Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *ICCV*, 1999, pp. 666–673.
3. Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *IROS*, vol. 3, 2004, pp. 2301–2306.
4. R. Unnikrishnan and M. Hebert, "Fast extrinsic calibration of a laser rangefinder to a camera," Robotics Institute, Pittsburgh, Tech. Rep. CMU-RI-TR-05-09, 2005.
5. D. Scaramuzza, A. Harati, and R. Siegwart, "Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes," in *IROS*, 2007, pp. 4164–4169.
6. S. Fuchs and G. Hirzinger, "Extrinsic and depth calibration of tof-cameras," in *CVPR*, 2008, pp. 1–6.
7. D. Lichti, "Self-calibration of a 3D range camera," *ISPRS*, vol. 37, no. 3, 2008.
8. J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of time-of-flight depth and stereo for high accuracy depth maps," in *CVPR*, June 2008, pp. 1–8.
9. "Gesture keyboarding," United State Patent US 2010/0 199 228 A1, Aug. 5, 2010.
10. J. Bouguet. (2010, Mar.) Camera calibration toolbox for matlab. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/
11. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.