EXTRACTING 3D FROM 2D: SELECTION BASIS FOR CAMERA CALIBRATION

Maria-Cruz Villa-Uriol	extsuperscript{1}, Gautam Chaudhary	extsuperscript{1}, Falko Kuester	extsuperscript{1}, Tara Hutchinson	extsuperscript{2} and Nader Bagherzadeh	extsuperscript{1}

	extsuperscript{1}Department of Electrical Engineering and Computer Science
	extsuperscript{2}Department of Civil and Environmental Engineering

University of California - Irvine
Irvine, California, USA
email: [mvillaur, gchaudha, fkuester, thutchin, nader]@uci.edu

ABSTRACT

The recovery of three-dimensional information from a set of two-dimensional images requires the use of an accurate, and fast calibration technique. Although diverse techniques have been developed for single and multiple camera configurations, selection of the most appropriate task-specific technique is a non-trivial exercise. Therefore, in this paper, we present a review of calibration techniques, their respective application domains and associated boundary conditions. Boundary conditions of particular importance include, desired capture volume, available time budget, required accuracy, number of cameras and type of feature detection. Commonly used calibration tools are discussed, including their specific features and target applications. A methodology to determine the suitability of a calibration tool for a specific setup and requirements using a synthetically generated environment, is presented. Finally, a multi-camera calibration tool is evaluated for its robustness in determining important calibration information under a number of envisioned camera configurations. The goal of this study is to aid the reader in the selection of the most suitable calibration techniques and tools appropriate for the task at hand.

KEY WORDS

Camera calibration, computer vision, scientific visualization.

1 Introduction

Applications ranging from entertainment to scientific visualization require characterization of objects and phenomena occurring in our three-dimensional (3D) world. A broad range of image capturing devices allow the collection of such 3D information. A fundamental problem that needs to be addressed is the recovery of depth information lost during projection. Although diverse techniques have been presented to solve this problem, a primary factor in the selection of a suitable approach to perform 3D recovery is the final application of this information. Additional aspects to consider include the constraints affecting the acquisition system, such as: (i) characteristics of the environment, (ii) capture volume, (iii) time budget, (iv) required accuracy and (v) desired speed.

Depending on the acquisition configuration, either monocular images or video sequences are used [1]. Alternatively, multiple static images or video sequences may be used. If an engineering length scale is required, during the calibration stage, information regarding the characteristics of the capturing system, camera locations and camera internal features must be incorporated [2][3].

The choice of a calibration technique or tool will determine relevant aspects in the posterior 3D points recovery, such as the achievable accuracy and frame rate. Depending on the camera lens, accounting for radial distortion (or not) is also decisive for accuracy. Sometimes only uncalibrated sequences are available and traditional calibration techniques can not be used, resulting in additional computational cost for the reconstruction [4][5].

The objective of this paper is to provide an overview of many of the available techniques and tools that can be used when a calibration stage is needed for a specific application. We intend to pinpoint the decisions that need to be made and that will affect the quality of the final outcome, critical in the use of images to describe 3D information. First, fundamental terminology is defined, followed by a classification of calibration methods from a user’s perspective and subsequently a generalization of the required steps to perform camera calibration. A subset of the most commonly available calibration packages is described, together with their advantages and disadvantages. Finally, a multi-camera calibration tool is evaluated for its robustness in determining important calibration information under a number of envisioned camera configurations.

2 Terminology

To reconstruct a scene from images, the relationship between the coordinates of a set of points in 3D space with the coordinates of their corresponding image points must be established. Equations for such a relationship are written in a world reference frame. To assist with defining these relationships and relating them with camera properties, camera’s characteristics are generally grouped into intrinsic (internal) and extrinsic (external) parameters, as shown in Figure 1.
The internal parameters describe how the camera forms an image. They offer a relationship between the image and the camera. These parameters are important for linking the pixel coordinates of an image point with the corresponding coordinates in the camera reference frame. They can be defined as the set of parameters needed to characterize the optical, geometric and digital characteristics of the given camera. The intrinsic attributes are focal lengths \( f, f_x, f_y \), pixel sizes \( s_x, s_y \) in the \( x \) and \( y \) directions, aspect ratio \( a \) computed as \( s_y / s_x \), principal point \( P \), image center \((o_x, o_y)\) and radial distortion coefficients \( k_1 \) and \( k_2 \). While \( f \) represents the distance between the center of projection and the retinal plane specifying the perspective projection, \( f_x \) and \( f_y \) are the lengths in effective horizontal and vertical pixel size units. \( P \) specifies the coordinates \((p_x, p_y)\) of the center of the lens, which is the piercing point of the camera’s coordinate frame \( z \)-axis with the camera’s sensor plane. Finally \( k_1 \) and \( k_2 \) describe the geometric distortion introduced by the optics of the lenses.

The extrinsic parameters define the camera’s position and orientation. These relate the camera to the real world, and are defined as any set of geometric parameters that uniquely identify the transformation between the unknown reference frame and a known reference frame, called the world reference frame. The extrinsic parameters are (a) 3D translation vector \( t \), providing the translational components for the transform between the world and camera coordinate frames, and (b) \( 3 \times 3 \) rotation matrix \( R \), a product of matrices \( R_x, R_y, R_z \) representing the rotation angles for the transform between the world and camera coordinate frames.

The principal point, effective horizontal length, aspect ratio and the image centers are required for the transformation between the frame coordinates and pixels coordinates. The task to estimate the values of the intrinsic and extrinsic parameters of a single camera or a set of multiple cameras is termed camera calibration.

### 3 Classification of Calibration Methods

Accuracy of the final recovered 3D dataset depends upon the accuracy of the calibration [6]. Ideally, the calibration tolerance should be one or two magnitudes smaller than the desired tolerance when 3D acquisition is conducted [2]. For this reason, ensuring an accurate calibration is essential and it directly derives from the qualities of the acquisition system, characteristics of the captured environment and the chosen calibration technique.

Strategic placement of cameras, acquired image resolution and quality of camera lens(es) affects the acquisition system in terms of space. At a temporal level, required acquisition frame rate impacts the amount of information that needs to be stored and analyzed to obtain an accurate calibration.

Several calibration techniques are widely used in different application domains. In Figure 2, we depict a classification of available calibration methods from an application user’s perspective. We organize them into two broad categories, based on the final usage and the importance of accuracy to these applications. In general, applications used for scientific and engineering purposes require significantly more accuracy than those in the entertainment industry. Thus, our focus will be on the scientific application domains, where a set of features such as color and shape guides the calibration process. We differentiate between systems that demand manual user intervention of features and those that are totally automated. Although manual identification of features simplifies the process, it is error-prone and may result in being more time consuming than the calibration process itself. Hence, automated or semi-automated approaches are desirable.

![Figure 1. Fundamental terminology in world and camera coordinate systems.](image)

![Figure 2. Classification of calibration methods.](image)
ible from most of the cameras. Features extracted from the images can be categorized very generally as either: (i) marker-based or (ii) pattern-based. Marker-based features may be either active or passive within the scene. Patterns may include multiple or single basic geometry primitives, however, generally include very simple and well-defined color transitions (e.g. black on white checkerboards).

4 Techniques and Steps Involved

Most scene reconstruction and machine vision projects treat calibration as a preprocess that involves several steps, including camera placement, data acquisition, feature detection and feature matching [7]. Figure 3 illustrates commonalities and discrete differences for the most common techniques. The calibration process generally begins with a careful definition of acquisition objectives, enabling the user to decide on the required acquisition area or volume as well as the needed spatial and temporal resolution.

4.1 Image Acquisition System Design

Single camera or multi-camera configurations can be used to generate the reference images needed for calibration. For a single camera, this may simply involve acquiring individual still images either from multiple viewpoints for a static scene, or from a fixed viewpoint for a time varying scene (i.e. a moving object). Camera arrays on the other hand require a mapping between features present in each of the available views of the scene. Using the corresponding points information, the implicit and explicit parameters of the cameras can be computed. Generally some image processing such as background substraction or filtering will be required to remove unwanted noise. All cameras need to be synchronized in order to ensure frame correspondence between cameras, making certain that a given frame number is captured at the same instant of time by every camera. For this type of a setup, synchronization is either done in software or in hardware and a high frames count from each camera is usually required for calibration [8]. Therefore, an automated processing sequence is desirable that facilitates time efficient processing of the reference images.

Passive or active markers can be introduced in the scene to simplify the identification of feature correspondences between images (Figure 4). This involves introducing a known pattern into the scene and to subsequently identify its features, such as known edges and colors to establish correspondence between different images. The reference markers can be useful beyond calibration, to accomplish object tracking itself, as illustrated in Figure 4.

4.2 Selection of Calibration Patterns

Camera calibration usually relies on the extraction of features from the available input images. A simple method to allow unit-based calibration, consists of having a reference object of known dimensions present in all the views. This reference simplifies the task of feature identification and allows for efficient calibration. The approach for single camera calibration uses planar objects with a very precisely known geometrical pattern. In this scenario, the calibration process and setup are relatively simple, as opposed to multiple camera calibration, where three dimensional cuboidal
5.1 Static Pattern Calibration

The calibration tools discussed in this subsection use a static, regular chessboard pattern as a reference. One example is the Bouguet camera calibration toolbox [15]. This toolbox was developed in C/C++, using Intel’s Open Source Computer Vision Library (OpenCV). This tool can calibrate a video camera accurately in a matter of just a few seconds. A flat checkerboard pattern in combination with knowledge about the dimensions of each square are used to enable automatic edge detection and computation of the intrinsic camera parameters (focal length, principal point, distortion coefficients) as well as the extrinsic parameters (3D position of the pattern for each image). Once calibration is complete, image distortion can be removed in real-time. The lens distortion model consists of two terms: a radial distortion term (up to the fourth order) and a tangential distortion term consisting of two scalars for encoding the angular orientation of the focal plane with respect to the sensor plane. Consequently, the lens distortion model is parameterized using four scalar coefficients. One drawback of this toolbox is that it only supports the calibration of one camera at any given time. A Matlab implementation is available as well but is limited by performance bottlenecks.

Heikkilä and Silven have developed another Matlab based open-source calibration tool [16][13][11]. The algorithm is based on a circular pattern and uses a new bias correction procedure for circular control points and a non-recursive method to reverse the distortion model. A drawback of this approach is that it does not support automatic corner detection, requiring that users specify values, identifying the corners of the checkered pattern. Zhang [17][18] developed a flexible camera calibration technique. However, it does not have corner extraction and hence, user input is required to provide this information. The Tsai camera calibration application is considered one of the classic tools and is based on Tsai’s algorithm for the perspective projection camera model [9][19]. The foundation for this algorithm is the pinhole model for perspective projection. This tool as well does not include an automatic corner extractor.

5.2 Dynamic Calibration - Single Camera

Dynamic calibration encompasses the scenario where either a single moving camera (or video camera) is used to collect a sequence of images over a period of time. Commonly termed video sequence calibration, this approach involves identifying a number of features that are shared between consecutive frames. These features are then tracked from frame to frame, and thus form the points of correspondence for calibration. A drawback of this method is that commonly, extended image sequences (hundreds of frames) are required. If the user must identify the different features in each frame, then the delay in the process is dominated by the user’s input time. Hence, an automated system is highly desirable in such video sequence calibration approaches and one example of such an application is

5 Common Calibration Packages

A number of common calibration packages exist, that mainly differ in the level of accuracy that can be obtained. Some of those are discussed in the following sections.
discussed in [1].

5.3 Multi-Camera Calibration
In this section, we describe selected multi-camera calibration application frameworks. One such package, developed within MatLab and C is the EasyCal [8] calibration software. The engine behind EasyCal is an extension of the camera calibration toolbox developed by Bouguet [15]. The main advantage of this tool is the modular multi-image feature, while a primary drawback is the large number of images required. The authors note that a sequence of at least 1000 images at a relatively high frame rate (≥ 15fps) are required for suitable results [8]. In addition, images of a basic chessboard design pattern from at least two cameras are needed. A second drawback is the approach expects the room lighting to be turned off while acquiring the point images as calibration data. This may not be practically possible for many applications.

The Multi-Camera Self-Calibration Toolbox [14], developed at the Computer Vision Laboratory of the Swiss Federal Institute of Technology, incorporates each calibration phase within the MatLab toolbox. A minimum of three cameras and a diffused laser pointer, which is tracked during calibration, are required for this algorithm. This software is well suited for calibrating multiple cameras at the same time. An important advantage of this toolbox is that it is completely automated. The light source is moved throughout the anticipated field of view during the calibration phase, generating a set of images of the sources’ path. The light source does not necessarily need to be visible from all cameras during the entire calibration capture, as the algorithm compensates for occluded or missing points. The only drawback of this application is that it requires a relatively dark surrounding, such that the light source maintains a good contrast with the background for tracking purposes.

6 Case Study of a Calibration Tool
To test the feasibility of the Multi-Camera Self-Calibration Toolbox [14], we evaluate its capabilities considering a variety of envisioned cameras and locations. The purpose of this work is to test if this tool is suitable for a given application, in terms of volume to be captured, achievable accuracy, robustness, storage needs and calibration time budget. These parameters are used as a basis to determine if the combination of camera setup and calibration tool studied meets certain requirements. The camera setup is defined by a set of intrinsic and extrinsic parameters (camera placement) of all the cameras in consideration. A general methodology is presented to test different pairs of camera setups and calibration tools to help in the decision process when designing a multiple camera capture system. Our approach utilizes a synthetically generated multi-camera capture setup recreating real world conditions to test the calibration tool, in combination with a synthetic calibration sequence.

A three-dimensional modeling tool was used to generate arbitrary scenes with a variable number of n cameras (3 to 12) with known intrinsic and extrinsic parameters. In this paper we present an analysis of 12 setups: (a) four setups of four cameras each, (b) two setups of eight cameras and (c) six different setups of 3, 5, 6, 7, 10 and 12 cameras each. A set of snapshots showing top and perspective views of a subset are included in Figure 6. More specifically, there is one case with n = 12 placed in two concentric circles at two different heights and at 30 degrees angular intervals (Figure 6 (c) and (d)). Two different cases with n = 4 are also included, corresponding to cases (ii) and (iii). Case (ii) shows an arbitrary placement of cameras, while in case (iii) they are placed in a single half of the hemisphere covering the scene.

The multi-camera self-calibration toolbox requires a calibration sequence in which each of the cameras captures a colored-light source moving through the capture volume. In our tests, each of these calibration sequences consisted of 300 frames with a resolution of 640x480 pixels, covering a capture volume of 10x10x10 units with a red-light source of radius 0.5 units. The viewing volumes for all studied camera configurations were of similar size.

To begin our study, the calibration accuracy of each camera setup needs to be determined. To perform this task, we analyzed the differences between the calibration results obtained from the calibration tool and the known, synthetically generated setup. Intrinsic and extrinsic camera parameters of each of the cameras are obtained and compared, namely, focal length (f_x, f_y) and camera location expressed as a rotation (rot) and a translation vector (trans). In the case of focal length the error between the calibrated
Figure 7. Computed errors for a 12 cameras setup.

\[ \text{error}\% (f_x) = \frac{f_{x, \text{cal}} - f_{x, \text{syn}}}{f_{x, \text{syn}}} \times 100 \]
\[ \text{error}\% (f_y) = \frac{f_{y, \text{cal}} - f_{y, \text{syn}}}{f_{y, \text{syn}}} \times 100 \]

where \( f_{x, \text{cal}} \) and \( f_{y, \text{cal}} \) are the computed values for \( f_x \) and \( f_y \) by the calibration tool, and \( f_{x, \text{syn}} \) and \( f_{y, \text{syn}} \) are the known values for \( f_x \) and \( f_y \) obtained from the synthetic scene setup.

The square error is used for camera locations,

\[ \text{square error (rot)} = \frac{1}{9} \sum_{i=1}^{3} (r_{\text{cal}_{ij}}^2 - r_{\text{syn}_{ij}}^2) \]
\[ \text{square error (trans)} = \frac{1}{3} \sum_{i=1}^{3} (t_{\text{cal}_i}^2 - t_{\text{syn}_i}^2) \]

where \( r_{\text{cal}_{ij}} \) correspond to the \( i \) by \( j \) elements of the rotation matrix \( R \) and \( t_{\text{cal}_i} \) to the translational components of the normalized vector \( \tilde{t} \) characterizing the transform between the world and camera coordinate frames, obtained from the calibration process. \( r_{\text{syn}_{ij}} \) and \( t_{\text{syn}_i} \) are the known values extracted from the synthetic scene.

The plots in Figure 7 show a sample of the computed errors for each of the individual cameras in the twelve camera setup shown in Figure 6(c) and (d). The impact each camera has on the computation of the mean error for the camera setup (\( \text{mean } f_x, \text{mean } f_y, \text{mean } \text{trans} \) and \( \text{mean } \text{rot} \)) are also noted. Results presented in Figure 7 illustrate the high accuracy obtained, in terms of focal length and translation and rotation matrices. The mean error (amongst all cameras in this setup) was very small 0.25%. In addition, the maximum error observed, in terms of focal length was 0.8%.

A comparison among all twelve tested setups in terms of the mean error for \( \text{mean } f_x \) and \( \text{mean } f_y \) and the mean square error for \( \text{mean trans} \) and \( \text{mean rot} \) are presented in Figure 8. These are presented as the mean of all cameras used in the setup considered. Among the four different setups with four cameras each, there are some differences in the observed error. For example, the mean focal length error ranges from 0.2% to 1.6% for cases (ii) and (iii). This behavior can be explained with the different camera placement used for each setup (Figure 6). Case (iii) of the four-camera setup was designed with all cameras located on the same side of the hemisphere, resulting in more overlap of the viewing volume. As one might expect, as the number of cameras increase in a setup, the computed mean error decreases.

It also can be observed that a high correlation exists in the error in focal lengths among the different setups, which is in the order of 0.4%. In addition, rotation and translation square errors are very small, with general magnitude below the range of 10^{-5}. Analysis of different camera configurations demonstrates that the expected calibration errors introduced by the calibration tool can be bounded. As such, this methodology provides a platform to test and compare
arbitrary calibration tools and camera setups.

7 Summary and Concluding Remarks

The objective of this paper is to provide an overview of many of the available techniques and tools that can be used when a calibration stage is needed to recover 3D information from a set of two-dimensional images. This is well recognized as a critical, yet tedious and non-trivial part of the processing pipeline. However, until now, a unified discussion of this topic, where the advantages and disadvantages of a variety of techniques are explored, has not been presented in the literature. Therefore, we first provide a classification of calibration methods from a user’s perspective and subsequently generalize the required steps to perform camera calibration. A subset of the most commonly available calibration packages are summarized. The goal of this literary review is to aid the reader in the selection of the most suitable calibration techniques and tools appropriate for the task at hand. Finally, a methodology to determine the suitability of a calibration tool for a specific setup and requirements, is presented. We conclude that with the tested tool, low errors for intrinsic and extrinsic camera properties may be obtained, for the range of camera configurations considered.

Acknowledgements

Support for this study was provided by the CUREE-Kajima Phase V research program. We would like to thank CUREE-Kajima and our project manager Dr. Kanda for their support.

References


