A SEMI-AUTOMATED FRAMEWORK FOR HOMOGRAPHY ESTIMATION

Thomas Fäulhammer
Automation and Control Institute
Vienna University of Technology
Güßhausstraße 27-29 / E376
1040 Vienna, Austria
email: thomas.faeulhammer@tuwien.ac.at

Paulo Vinicius Koerich Borges *
Autonomous Systems Laboratory
CSIRO ICT Centre
1 Technology Ct
Pullenvale QLD 4069, Australia
email: Paulo.Borges@csiro.au

ABSTRACT
Reliable estimation of the geometric transformation (homography) between different planes is a key step in a number of robotics applications. A common requirement is the homography estimation between a ground plane and a camera. For this task, the most usual approach is to define the correspondence between the coordinates of real-world points and their pixel coordinates in the image. Often, however, real-world coordinates are not available or ground measurements are not practical. Targeting these scenarios, a homography estimation framework is proposed using (semi-)rigid objects that can be placed in different positions in the field of view, with its top and bottom pixels approximately identified in the image. The derived method shows that it can work with or without the knowledge of the camera intrinsics, even in presence of non-Gaussian measurement noise. A thorough analysis of the method is presented by observing the effect of noise, distance from the camera and number of measurements on the quality of the final results. The experiments illustrate the applicability of the method.

KEY WORDS
Camera Calibration - Homography Estimation - 3D Projection.

1 Introduction

The geometric transformation that defines the correspondence between points in different planes is widely known as projective transformation or homography. In computer vision and robotics, the homography finds applications in localization, recognition, registration, reconstruction and feature-based object detection methods [1, 2].

A simple way to calculate the homography is to find points on the image plane and their corresponding points on the area of interest in the real world. This requires the knowledge of world coordinates and their respective locations on the image plane [3]. In situations where the real world coordinates are not known in advance, a common way to obtain the homography is to exploit architecture and structured information on the projected image [2–6]. Instead of relying on structural information or world coordinates, this paper presents a framework that estimates the homography in a sparse environment. The required input for the homography calculation is obtained by placing an object at least four different positions on the surface of interest or by tracking an object moving around the surface, such as vehicles, humans or mobile robots.

To illustrate the applicability of the proposed method, experiments are performed on different target objects in situations with and without knowledge of the camera intrinsic parameters. The discussion addresses how results are affected by practical issues, such as the precision in finding the top and bottom pixels of the target object, the distance from the camera to the object and the number of measurements taken into account. Despite the fact that top and bottom points are not precisely detected, it is shown that coarse tracking of a semi-morphable object, such as a high visibility vest shown in Figure 12, over a significant number of frames still yields satisfying results.

This paper is organized as follows. Section 2 discusses related work and indicates how the present work differs from existing approaches. Section 3 formalizes the geometry problem involved in the homography estimation, serving as background for the subsequent sections. Section 4 presents the proposed methodology for homography estimation without point correspondence between the ground plane and the image pixels. A number of experiments are discussed in Section 5 followed by relevant conclusions in Section 6.

2 Related Work

To obtain the intrinsic parameters of the camera, a usual approach is to use an algorithm which automatically detects the corners of a checker board with sub-pixel accuracy [7]. In some environments, e.g. where the camera is too far away, this approach is not feasible due to the low resolution of small, distant objects. This is especially the case in surveillance situations, where the cameras are already installed. Other approaches use vanishing points and a vanishing line from sets of parallel lines on a projective plane [8–10]. To obtain good results, the space has to fulfill strong rigidity constraints of parallelism and orthogonality
that are in general not present in large, open and flat outdoor scenes.

Camera calibration by methods using information from tracked humans in the scene often require multiple cameras and a precise feature matching (finding foot and head location) of the human in the different camera views [11–13]. Approaches designed for single-views usually demand an accurate detection of a blob (connected components extracted from moving foreground), which is especially difficult when the object throws shadows or has reflections. Micsuk and Pajdla [14] proposed a foot-head homology estimation, which gives higher resolution of the detector by running a contour detector in two stages and using 3D projected models from changing a virtual camera viewpoint. Moving humans or objects, whose main axis can be uniquely identified in every possible view, are not always present in a scene, such as in industrial environments with vehicles or automated robots. In those cases, the addition of more geometrical information to the scene is necessary.

Liebowitz and Zisserman [4] used the decomposition of the homography matrix into similarity, affine and ‘pure projective’ transformation. The pure projective part can be determined by the vanishing line of the plane. The identification of the affine transformation requires two independent constraints with metric information of the environment, which can be a known angle, two equally though unknown angles or a known length ratio. Even though there are methods to calculate the angle between two line directions from their vanishing points and the image of the absolute conic in the image [3], it is not always possible to find enough constraints in a sparse environment. Using a similar approach, other works [2, 5] extended the set of possible constraints by tracking two objects that move with constant speed along two non-parallel linear paths. While in urban environments these constraints are often met, in an industrial site people tend to move more randomly in different directions and change their walking speed according to the surroundings. Another problem occurs when objects are moving along one line only. Chen et al. [6] proposed a conic based planar rectification method that relies on the presence of two (intersecting) circles.

Rather than relying on structural information of the scene, the novelty of the proposed approach is that the homography can be estimated by any objects having the same height in the real world (measured from the surface) and appearing at at least four different locations (possibly in different frames) in the image. In order to fulfill these requirements in scenes where these objects are not naturally present, it is possible to add arbitrary objects to the scene, e.g. traffic cones, beacons, persons or robots. In contrast to previously discussed methods adding structural information permanently to the scene, these temporary objects can be removed after successfully estimating the homography and thus do not change the environment in a larger scale. Using approaches discussed in this section, combining relevant methods and adjusting it to fit accordingly into the problem, the following procedure explains a simple solution for homography estimation, which can be obtained even in a sparse environment.

3 Background

To obtain a map, or so-called “bird’s-eye view”, from a captured image, the local relationship between points on the image plane and points on a surface plane must be determined. This relationship is described by the term homography. The remaining of this section provides a brief overview of the projective transform and defines basic notations used in this paper.

3.1 Camera Model

For simplification, the camera is described as pinhole model. The camera’s coordinate system is represented by coordinates \((x, y, z)\) and the world coordinate system by \((x', y', z')\). The focal length \(f\) is the shortest distance between the image plane and the camera centre \(c\) which also defines the origin of the camera coordinate system. The \(z\) axis is perpendicular to the image plane and goes through the principal point \(pp\) with the coordinates \((0, 0, f)\). The orthogonal axes \(x\) and \(y\) represent the horizontal and vertical direction respectively.

3.2 Homography

Points on a surface \(\pi\) in the physical world represented in the world coordinate system \((x', y') \in \mathbb{R}^2\) are imaged by a projective transform on the projection screen of the camera with coordinates \((x, y) \in \mathbb{R}^2\). Using homogeneous coordinates and the characteristics of the projective space, it is possible to find a \(3 \times 3\) matrix \(H\) that describes a linear relationship between these two coordinate systems up to a scale \(\lambda \in \mathbb{R}\setminus\{0\}::\)

\[
\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = H \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}.
\] (1)

Since the homogeneous matrix \(H\) is only defined up to a scale \((H_{33} = 1)\), it has 8 degrees of freedom and can be obtained by four different pairs of 2D coordinates \((x'_i, y'_i) \to (x_i, y_i)\). The restriction is that no 3 points can be collinear. In general, more pairs reduce the influence of noise present in the measurement. \(H\) is then calculated using all the point pairs minimizing a particular cost function, such as the forward-projection error \(\sum_i x_i \times Hx'_i\) or the back-projection error

\[
\sum_i (x_i - \frac{h_{11}x'_i + h_{12}y'_i + h_{13}}{h_{31}x'_i + h_{32}y'_i + h_{33}})^2 + (y_i - \frac{h_{21}x'_i + h_{22}y'_i + h_{23}}{h_{31}x'_i + h_{32}y'_i + h_{33}})^2.
\] (2)
3.3 Vanishing Line

Due to perspective transformation, a set of lines which are parallel in the real world intersect at a vanishing point \( \mathbf{v}_p \) on the image plane. Vanishing points of lines that are parallel to a plane \( \pi \) lie on a vanishing line \( \mathbf{v}_l \). Points on \( \mathbf{v}_l \) fulfill the equation

\[
\mathbf{v}_l \in \mathbb{R}^3 : ax + by + cz = 0,
\]

and can be obtained by the cross product of two vanishing points \( \mathbf{v}_{p1} \) and \( \mathbf{v}_{p2} \)

\[
\mathbf{v}_l = \mathbf{v}_{p1} \times \mathbf{v}_{p2}.
\]

The vanishing line \( \mathbf{v}_l \) resulting from the perspective transform provides information about the geometry of the surface plane with respect to the camera setup and is used in the following section to estimate the homography.

4 Approach

This section describes the gist of the proposed approach for homography estimation by considering two scenarios, one with pre-defined camera intrinsics and one in which this information is not available.

4.1 Using Camera Calibration Matrix

According to Hartley and Zisserman [3], the orientation of the surface plane is given by

\[
\mathbf{n} = \mathbf{K}^{-T} \mathbf{v}_l,
\]

where \( \mathbf{n} \) is the normal direction of the surface plane \( \pi \) and \( \mathbf{K} \) the intrinsic matrix of the camera. In an environment without sufficient structured information and architecture, it is often impossible to find two or more sets of parallel lines, which are required to calculate the vanishing line of the projected surface plane. To add necessary geometric information in the field of view, objects, which have the same shape (i.e. same height \( L \)) and are easy to identify on the captured images, are placed on the surface. If these objects are standing parallel to the surface normal \( \mathbf{n} \), they form sets of parallel lines, all parallel to the surface. Under perspective transform, the vanishing points of the objects lie on the same vanishing line \( \mathbf{v}_l \). To calculate this line, at least four objects in the field of the camera view are required (not necessarily in the same frame). Due to noise that occurs during the assignment of the bottom and top points of the object in the image, the vanishing line of the surface plane \( \pi \) can be defined as the straight line that minimizes the distance to the individual computed vanishing points. For this purpose the linear least square method is applied to Equation (3). In general, the estimation of the vanishing line improves with the number of objects but also depends on the location of the objects in the scene.

The following calculations assume a camera with square pixel, which is usually a good approximation for most modern camera chips. In this case, \( f_x \) and \( f_y \) are the same and further described as focal length \( f \).

Figure 1 shows the screen projection \( L_p \) of the object and the \( z \) component \( d_z \) of the location of the object in the camera coordinate system

\[
d_z = f \frac{L}{L_p}.
\]

As the estimation is based on a distance measurement that is replicable with a direct measurement on the surface plane \( \pi \) rather than a distance measurement along the camera axis \( z \), the distance \( d_z \) is projected to the real-world coordinate system as shown in Figure 2. This system is aligned to the camera system such that the \( z' \)-axis is rotated by the angle \( \psi \) with respect to \( z \), the \( y' \)-axis is parallel to the normal direction \( \mathbf{n} \) of the surface plane \( \pi \), and the axes \( x' \) and \( z' \) spanning a plane coplanar to the surface plane with the origin at the camera’s centre point \( \mathbf{c} \). The distance \( d_{z'} \) from the camera centre to the object along the \( z' \)-axis is given by

\[
d_{z'} = d_z \cos \psi.
\]
Similarly, the vertical position \( d'x \) in real-world coordinates is estimated by the projective geometry depicted in Figure 3

\[
d'x = \frac{x}{f} \cdot \cos \varphi, \tag{8}
\]

with the angle \( \varphi \) between \( x \) and \( x' \) axes obtained from the vanishing line \( v \).

For a given focal length and camera view, the projected height \( L_p \) varies according to

\[
\frac{\partial L_p}{\partial z} = -\frac{fL}{z^2}, \tag{9}
\]

\[
\Delta L_p = -fL \int_{z_1}^{z_2} \frac{dz}{z} = f \cdot L \cdot \left( \frac{1}{z_1} - \frac{1}{z_2} \right), \tag{10}
\]

with \( z_1 \) describing the closest and \( z_2 \) the most distant point to the camera.

Since the variable \( L_p \) carries information about the position of the object, the signal-to-noise ratio depends on the ratio of \( \Delta L_p \) to the variance of the measurement noise \( \text{Var}\{L_p\} \). This ratio influences the number of objects/frames needed to stay below a pre-defined accepted error of the homography estimation and depends on the camera environment and the accuracy of the object measurement (quality of detection - either by user or automated by a tracker).

All measurement errors of this approach can be reduced to projection errors of the height of the object. The influence of a noisy measurement of the projected height, expressed as the variance of this variable \( \text{Var}\{L_p\} \), yields by partial derivative of Equations (7) and (8) the variance \( \text{Var}\{d'x\} \) and \( \text{Var}\{d'y\} \) of the estimated \( d'x \)- and \( d'y \)-components:

\[
\text{Var}\{d'x\} = fL \left( \frac{\cos \Psi(L_p)}{L_p^2} + \frac{\sin \Psi(L_p) \cdot \partial \Psi(L_p)}{L_p} \right) \text{Var}\{L_p\}, \tag{11}
\]

\[
\text{Var}\{d'y\} = \frac{x}{f} \left( \frac{L}{L_p^2 \cos \phi(L_p)} + d'y \frac{\sin \phi(L_p) \cdot \partial \phi(L_p)}{\cos^2 \phi(L_p)} \right) \text{Var}\{L_p\}. \tag{12}
\]

If the vanishing line of surface plane aligns approximately with the \( x \)-axis of the camera, which is often the case in outdoor surveillance camera systems, these equations can be approximated by:

\[
\text{Var}\{d'x\} = \frac{fL}{L_p^2} \text{Var}\{L_p\}, \tag{13}
\]

\[
\text{Var}\{d'y\} = \frac{xl}{L_p^2} \text{Var}\{L_p\}. \tag{14}
\]

Therefore, the influence of an error of the object detection on the estimated world coordinates is dependent on the location of the object in the scene and can in general be reduced by maximizing the height of the object in the image.

### 4.2 Unknown Calibration Matrix

In certain scenarios, the calibration matrix \( K \) is not known and difficult to obtain. If due to the lack of a calibration matrix a slightly higher tolerance in the mapping error is accepted, the measurement can be approximated by the movement of a semi-rigid object within a single camera view. Neglecting a possible skew factor and assuming that the principal point \( pp \) is in the middle of the image (cp. [8]), the focal length \( f \) is the only parameter that has to be provided. According to Lai and Yilmaz [18], the focal length can be estimated using the information of the vanishing line \( v_1 \) of the surface plane \( \pi \) and a vanishing point \( v_z \) of a reference line perpendicular to the plane \( \pi \). The point \( p \) is the intersection point of the vanishing line \( v_1 \) and a line perpendicular to \( v_1 \) that goes through the vertical vanishing point \( v_z \) such that

\[
0 = v_1 \cdot (pp \times v_z), \tag{15}
\]

\[
p = v_1 \times (pp \times v_z). \tag{16}
\]

Neglecting any lens distortions, \( p, pp, v_z \) and \( c \) are the corner points of two triangles depicted in Figure 4. The focal length \( f \) is the distance between \( c \) and \( pp \). Writing the distance between two points \( p_1 \) and \( p_2 \) as \( d(p_1, p_2) \), the focal length can be written as

\[
f^2 = d(p, pp) \cdot d(pp, v_z). \tag{17}
\]
Due to noise in the process, the vertical lines often do not intersect at the same point and therefore the vertical vanishing point \( v_z \) is not well defined. Note that the computation of \( v_z \) by intersecting the lines pairwise, removing outliers and using the centroid of these intersections is not an optimal procedure.

Under the assumption of Gaussian measurement noise, the maximum likelihood estimate (MLE) of the vanishing point and line segments is computed by determining a set of lines that do intersect in a single point and minimize the sum of squared orthogonal distances from the endpoints of the measured line segments. This minimization may be computed numerically using the Levenberg-Marquardt algorithm.

Under degenerated conditions, where the image plane is orthogonal (front view) or parallel (top view) to the surface plane \( \pi \), either the vertical vanishing point \( v_z \) or \( p \) falls at infinity. Clearly, the computation fails under this condition or, in the case of a top view, makes any processing unnecessary.

5 Results

In order to verify the proposed method, this section discusses experiments in an industrial environment using two different camera views with slightly different focal lengths and principal points. For each view, two different approaches are employed for the height measurement of the object.

The following results were obtained with the camera AXIS Q6034-E (resolution: \( 1280 \times 720 \) pixels) mounted about 10 m above the surface plane \( \pi \). For validation, the ground truth is taken from previous measured real-world coordinates of image points. Figures 6 and 12 show the environment of camera view 1 and 2, respectively. Some of the ground truth points are depicted as letters A-I in the images and describe locations of obstacles for autonomous vehicles. One of the goals is to estimate the real world distance to the closest obstacle by estimating the homography only with a single camera view.

Since the proposed method can only measure length ratios and angles, the results are scaled and rotated to fit the ground truth locations prior to computing a validation error.

With respect to identifying the points of interest in the target object, the top and bottom points of the object were selected manually by the user. Alternatively, the object is tracked by simple color thresholding combined with foreground detection and morphological transformations. To calculate the homography of the image, the algorithm uses the location and height of the object in each frame the object has been detected. The focal length is once taken from the intrinsics matrix of a pre-calibrated camera and once computed automatically by the approach described by Lv et al. [11] and reviewed in Section 4.2. In the latter case, the intrinsics are not known in advance.

5.1 Homography Estimation of an Industrial Environment

The input data of the experiment is given by a yellow painted pipe shown in Figure 5 with a height of 1.12 m and a distinctive color which is uncommon in the test environment for easy detection and tracking (cp. [19]). Once the object is placed manually on different locations in the camera view and the corresponding image points are selected by hand, and once the object is mounted on a platform of a slowly-moving mobile robot and the image points tracked automatically. Due to the rigidity of the object and a careful selection of the object points in the image, removing potential outliers, the assumption of Gaussian measurement noise is approximately fulfilled.

The focal length in this experiment is 2000 px (square pixel assumption verified as reasonable) with the closest point and most distant point in this view approximately 30 m and 60 m away from the camera. According to \( (10) \), the variation of the height of the object in the captured image is \( \Delta h_p \approx 2000 \text{ px} \cdot 1.12 \text{ m} \cdot \left( \frac{1}{30 \text{ m}} - \frac{1}{60 \text{ m}} \right) = 37.3 \text{ px} \).

The mapped obstacles A-I with the matrix \( H \) for View 1 and View 2 are depicted in Figure 7 and 10, respec-
tively. The resulting error of the homography estimation is shown in Table 1.

The transformation of the whole image can also be used as validation for the estimated homography. Figure 8 shows the warped picture reflecting known real-world properties of a grid (perpendicular lines and cell size \(3 \times 1\)) on the measured surface \(\pi\).

Figure 9 indicates the influence of the number of measurements on the average Euclidean error of 15 ground truth locations and various object selection methods. Considering a significant number of measurements, the experiment demonstrates that the quality of the estimation can be reasonably accurate even with a simple tracker based on colour and blob detection. Given that there are more than about 80 frames taken into consideration, the experiment shows that the automated approach provides the same quality as the approach where the user has to select the points manually. The experiment further shows that good results can be achieved by estimating the focal length without the knowledge of the intrinsics of the camera as described in Section 4.2.

The influence of the signal-to-noise ratio of the height measurement on the error of the homography transform is depicted in Figure 11. The results are obtained by simulating additive Gaussian noise with variance \(\text{Var}\{L_p\}\) and demonstrate the importance of an accurate height measurement.

Figure 12: Effects of the homography transform on length ratios and angles when applying the estimated matrix \(H\) on the original image (a) to obtain a top view (b). To verify the quality of homography, the properties of the grid on the ground shown in red is compared to the ground truth of the grid (\(3 \times 1\) and \(90^\circ\)).

### 5.2 Effects of Non-Gaussian Measurement Noise in the Homography Estimation Process

Instead of tracking a distinctive pipe, this experiments tracks a semi-rigid high visibility vest worn by a person walking around the viewed area as shown in Figure 12. Due to the noise and outliers present in the detection, it was observed that it is difficult to obtain a good mapping with a small average re-projection error when all measurements are included in computing the cost function (2). Depending on the tracking parameters, the automated object detection tends to either under- or overestimate the height \(L_p\), which results in noise with non-Gaussian distribution. Since outliers decrease the accuracy of this homography estimation, considering all points to compute the homography estimate (regular method) does not decrease the average error with an increasing number of samples as shown in Figure 13.

After a reasonable amount of samples, a significantly smaller error can be obtained by using only a proper selection of points defined by one of the subset-based methods, RANSAC or LMedS (Table 1 and Figure 10).

There are different aspects leading to systematic errors in the measurement. As there even is a small error when computing the homography by the real-world coor-
Figure 9: Average Euclidean error of 15 validation points between ground truth real-world locations and mapped image points using the estimated homography, and scaling and rotating the mapped points accordingly. It shows the quality of the homography estimated by manual selection and using the calibration matrix (blue, solid); manual selection and estimating the focal length (red, dotted); and automatic detection of object using the intrinsics (green, dashed). The homography of the automated approach was calculated by RANSAC. For comparison, the error of the raw data (without homography transformation) is shown in solid magenta.

dinates [3] indicated in Table 1, it is expected that the surface plane is not completely flat. Furthermore, the main axis of the object can have a small tilt with respect to the surface normal $n$, which means the beacons are not standing perfectly straight. Another source of error comes from assigning top and bottom points of the objects, which is difficult due to the width of the objects, the finite resolution of the image, occlusions and bad lighting conditions (e.g. in the shade).

6 Conclusion

This paper presented a semi-automated framework for ground plane to camera homography estimation. The method does not require knowledge about real-world coordinates of points on the ground plane. The top and bottom points (pixel coordinates) of the target object are identified in the image with the object standing at different positions in different frames around the area of interest.

Depending on the method used for the selection process, the error of the Euclidean distance measurement compared to raw image evaluation was decreased to about 5-20% on average. The results indicated that even with low accuracy in detecting the top and bottom points, the methodology can robustly compute the homography estimation.

The proposed approach works especially well with wide-angle cameras which are used in most safety surveillance setups but requires a high-quality detection of the objects when using lenses with a relatively long focal length with respect to the distance between object and camera.

It finds particular applicability in areas where measuring real-world coordinates is hard or in non-permanent camera setups, where a fast and reliable homography estimation is necessary.
Figure 12: Estimating the homography of View 2 by tracking a high visibility vest.

Figure 13: Average Euclidean error of 15 validation points between ground truth real-world location and mapped image points using the estimated homography of the automated approach, and scaling and rotating the mapped points accordingly. It shows the quality of the homography, once calculated regularly by considering all measurements (blue, solid), and once by only considering a certain subset of the measurements defined by the methods LMedS (green, dashed) and RANSAC (red, dotted). For comparison, the error of the raw data (without homography transformation) is shown in solid magenta.

Acknowledgements

This work is supported by the CSIRO Minerals Down Under Flagship and the CSIRO ICT Centre.

References


