This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Minimal Solutions for Panoramic Stitching Given Gravity Prior

Yaqing Ding[1], Daniel Barath[2], Zuzana Kukelova[3]
[1] School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] Computer Vision and Geometry Group, Department of Computer Science, ETH Zürich
[3] Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague
dingyaqing@njust.edu.cn

Abstract

When capturing panoramas, people tend to align their cameras with the vertical axis, i.e., the direction of gravity. Moreover, modern devices, e.g. smartphones and tablets, are equipped with an IMU (Inertial Measurement Unit) that can measure the gravity vector accurately. Using this prior, the y-axes of the cameras can be aligned or assumed to be already aligned, reducing the relative orientation to 1-DOF (degree of freedom). Exploiting this assumption, we propose new minimal solutions to panoramic stitching of images taken by cameras with coinciding optical centers, i.e. undergoing pure rotation. We consider six practical camera configurations, from fully calibrated ones up to a camera with unknown fixed or varying focal length and with or without radial distortion. The solvers are tested both on synthetic scenes, on more than 500k real image pairs from the Sun360 dataset, and from scenes captured by us using two smartphones equipped with IMUs. The new solvers have similar or better accuracy than the state-of-the-art ones and outperform them in terms of processing time.

1. Introduction

Panoramic image stitching is a fundamental problem in computer vision. It is useful not only for creating visually pleasing image mosaics, but also for combining inputs from multiple cameras before further processing in downstream tasks. The generated images cover a large field-of-view and are useful in various applications, *e.g.* image-based localization [3], SLAM [25, 18], autonomous driving [33], sports broadcasting [7], video surveillance [35], and augmented and virtual reality [17, 27].

When solving this problem, we are given a sequence of images taken from a single point in space with a camera rotating around some 3D axis. The objective is to map the images into a common reference frame and to create a larger image composed of the captured ones, thus, covering a much wider field-of-view than each individual image. In other words, the goal is to estimate the unknown relative rotation and intrinsic parameters of cameras with coinciding optical centers, *i.e.*, cameras undergoing a pure rotational motion. This problem is often considered to be solved in the computer vision community, with a number of existing solutions [15, 5, 10, 19, 6, 22]. However, in this paper, we will show that the existing solutions do not exploit all available information that can be easily obtained from recent devices. This information can be used to simplify the problem formulation and to speed up the solution.

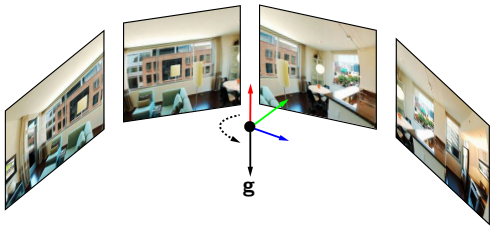
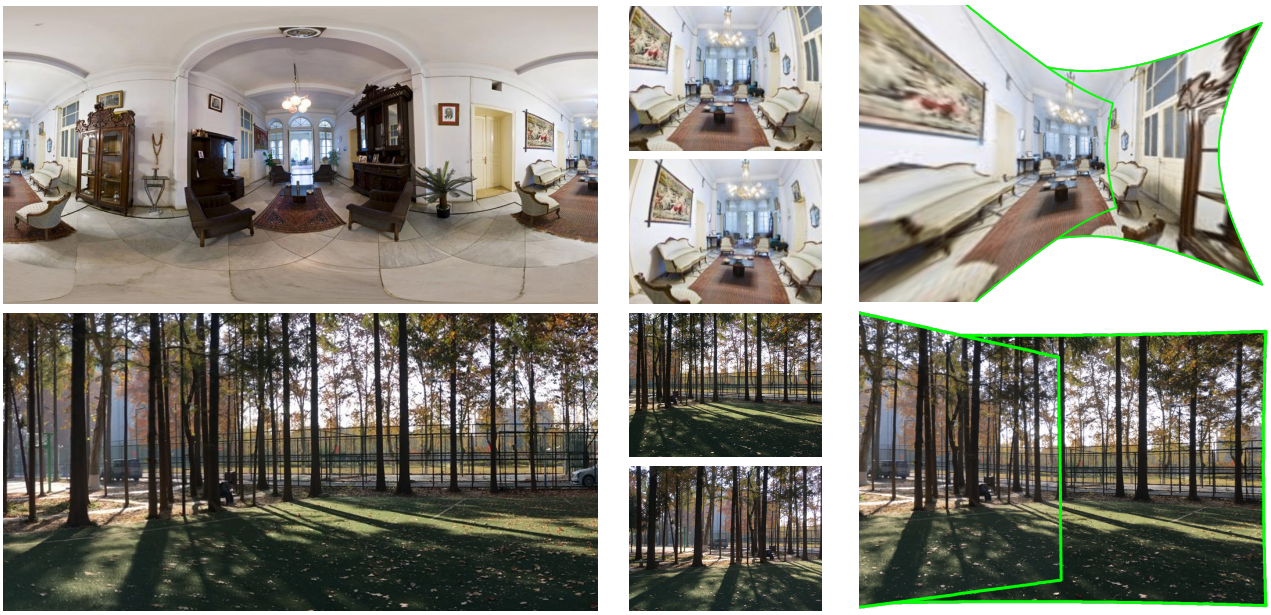


Figure 1: Panorama stitching with known gravity vector $\mathbf{g}$.

A typical panorama stitching pipeline consists of the following three major steps.

1. *Pair-wise matching*: In the first step, image features are obtained and tentative correspondences are matched between all image pairs.
2. *Robust pair-wise estimation*: From correspondences found in the first step, the intrinsic and extrinsic parameters of pairs of cameras are estimated robustly. Also, the correspondences consistent with the parameters are selected. The robust estimation is usually based on solving the stitching problem from a minimal number of input correspondences, *i.e.*, solving the problem using a minimal solver, in a RANSAC framework [5].
3. *Bundle adjustment (BA)*: The pair-wise estimates of intrinsic and extrinsic parameters are refined jointly over all images using a non-linear optimization.

In this paper, we address the second step of this pipeline, *i.e.*, the robust pair-wise estimation from a minimal number of correspondences. Several minimal solvers exists that can



(a) Full panoramic image (b) Image pair (c) Stitching

Figure 2: Example scenes and stitching results from the SUN360 (top) and our Smartphone (bottom) datasets.

be used in this step of the panoramic stitching pipeline. One of the basic well-known solvers for stitching assumes fully calibrated pinhole cameras and estimates the unknown rotation from two point correspondences [16]. However, the camera intrinsic parameters may not be known in practice, which is often the case in the real-world. Solutions to different camera configurations can be split into two main categories: the unknown focal length for narrow field-of-view cameras, and the unknown focal length and radial distortion case for wide-angle cameras. For the unknown focal length case, one of the most commonly used solvers is the normalized 4 point linear solution for homography estimation [15]. In [5], the authors demonstrated that 2 and 3 correspondences are sufficient for obtaining the homography induced by a rotation with 1 and 2 unknown focal lengths.

In practice, almost all cameras exhibit some amount of lens distortion. Moreover, wide-angle lenses with significant radial distortion are now commonly used in smart devices. These wide-angle cameras introduce large distortions that cannot be ignored. Previous work [10, 19] showed that modeling lens distortion inside the minimal solver is critical for obtaining high-quality homography estimates. Without modeling distortion in the pair-wise step, not only the initial pose and calibration estimates are less accurate but the set of inliers used for the non-linear optimization does not cover the whole image and misses highly distorted points from around the borders. Without such points, the non-linear optimization in the BA stage may not be able to correctly estimate the lens distortion and to produce good stitching results. To address this problem, Fitzgibbon [10] used the single parameter division model to simultaneously estimate the homography and radial distortion using 5 point correspondences. With this division model, a minimal 3 point solver was proposed for panoramic stitching with equal and unknown radial distortion [19, 6]. In [22], two solvers were proposed for estimating the homography between two cameras with different radial distortions.

All the mentioned minimal solvers estimate the full 3-DOF rotation. Especially for cameras with radial distortion, this makes the resulting system of polynomial equations complicated. Panoramas are usually captured with cameras or mobile phones *aligned* with the direction of gravity. Moreover, recent devices usually are equipped with an IMU sensor that can measure the gravity direction accurately. Using the gravity prior, the y-axes of the cameras can be aligned, reducing their relative orientation to 1-DOF. This prior not only simplifies the geometry and polynomial systems that have to be solved but, also, reduces that number of correspondences needed for the estimation. This is extremely important since the run-time of RANSAC-like robust estimation depends *exponentially* on the sample size. The gravity prior was used to simplify minimal relative pose [11, 12, 28, 32, 24, 30, 8, 9], absolute pose [21, 31, 2], and general radial distortion homography solvers [29]. Surprisingly, it was not considered in panoramic stitching.

We present the first minimal solutions to panoramic stitching of images taken by two cameras with coinciding optical centers, *i.e.*, undergoing pure rotation, exploiting a gravity prior. We consider six practical configurations:

(i) $\mathbf{H1(G)}$ - **Calibrated case**: The images are captured

by a camera with known focal length (*e.g.*, from EXIF-tag) and known or negligible radial distortion.

(ii) $\mathbf{H1f(G)}$ - **Equal focal length**: The images are captured by a camera with fixed unknown focal length and known or negligible radial distortion.

(iii) $\mathbf{H1\lambda(G)}$ - **Equal radial distortion**: The images are captured by a camera with known focal length (e.g., from EXIF-tag) and fixed unknown radial distortion.

(iv) $\mathbf{H2\lambda f(G)}$ - **Equal focal length and radial distortion**: The images are captured by a camera with fixed unknown focal length and radial distortion.

(v) $\mathbf{H2f_{1,2}(G)}$ - **Varying focal lengths**: The images are captured by cameras with different focal lengths, *e.g.*, a zoom camera, with negligible or known distortion.

(vi) $\mathbf{H3\lambda_{1,2}f_{1,2}(G)}$ - **Varying focal lengths and radial distortion**: The most general case, where the images are captured by cameras with different focal lengths and distortions, *e.g.*, using a wide-angle camera that was zooming during the image capture.

Experiments both on synthetic data and more than 500k real image pairs demonstrate that our new solvers are superior to the state-of-the-art methods in terms of processing time while leading to comparable or better accuracy.

Note, that the application of our solvers is not only limited to panoramic stitching. Such solvers can also be used in other applications, *e.g.*, for calibrating rotating surveillance [1] or PTZ cameras in sport matches [7], where the cameras undergo purely rotational motion, and to generate images with larger field of view for subsequent tasks, *e.g.* localization, and as inputs to CNNs.

2. Problem Statement

Given a set of 3D points $\{\mathbf{X}_i\}$ observed by two cameras undergoing a pure rotational motion, let $\mathbf{p}^d_{1i} = [u^d_{1i}, v^d_{1i}, 1]^\top$ and $\mathbf{p}^d_{2i} = [u^d_{2i}, v^d_{2i}, 1]^\top$ be the corresponding measured distorted image projections of $\mathbf{X}_i$ in these two cameras in their homogeneous form. Undistorted image points $\mathbf{p}^u_{1i} = u(\mathbf{p}^d_{1i}, \lambda_1)$ and $\mathbf{p}^u_{2i} = u(\mathbf{p}^d_{2i}, \lambda_2)$, undistorted with some undistortion function $u(\cdot, \boldsymbol{\lambda})$, are related by

$$d_{2i}\mathbf{K}_2^{-1}\mathbf{p}^u_{2i} = d_{1i}\mathbf{R}\mathbf{K}_1^{-1}\mathbf{p}^u_{1i}, \tag{1}$$

where d_{1i}, d_{2i} are the projective depths, $\mathbf{K}_1, \mathbf{K}_2$ are the intrinsic camera matrices, and $\mathbf{R} \in \mathrm{SO}(3)$ is the unknown relative rotation between the cameras.

We use the one-parameter division model [10] to parameterize the radial undistortion function u. This model is especially suited for minimal solvers since it is able to express a wide range of distortions with a single parameter and usually results in simpler equations compared to other distortion models. This model has been used for the simultaneous estimation of two-view geometry and lens distortion [10], for the estimation of radial distortion homography [22], and, also, for panorama stitching with radial distortion [6, 19].

In the one-parameter division model [10], the undistortion function u, that undistorts an image point with homogeneous coordinates $\mathbf{p}^d = [u^d, v^d, 1]$ using undistortion parameter λ, has the following form

$$\mathbf{p}^u = u(\mathbf{p}^d, \lambda) = [u_d, v_d, 1 + \lambda(u_d^2 + v_d^2)]^\top. \tag{2}$$

In this paper, we assume that the gravity direction is known. This is a reasonable assumption since panoramas usually are captured by cameras or mobile phones aligned with the direction of gravity. Moreover, smart devices are equipped with IMU sensors that can measure the gravity vector accurately. Using the gravity direction, we can compute the roll and pitch angles and use them to align the y-axes of the cameras. Let us denote the known rotation matrices used for the alignment of the two cameras as $\mathbf{R}_1$ and $\mathbf{R}_2$. In this case, (1) can be written as

$$d_{2i}\mathbf{R}_2\mathbf{K}_2^{-1}\mathbf{p}^u_{2i} = d_{1i}\mathbf{R}_y\mathbf{R}_1\mathbf{K}_1^{-1}\mathbf{p}^u_{1i}, \tag{3}$$

where $\mathbf{R}_y$ is the unknown rotation from the yaw angle (the unknown rotation around the y-axis).

For most modern CCD or CMOS cameras, it is reasonable to assume that the camera has square-shaped pixels and that the principal point coincides with the image center [14]. In this case, the calibration matrices have the form $\mathbf{K}_1 = \mathrm{diag}(f_1, f_1, 1)$ and $\mathbf{K}_2 = \mathrm{diag}(f_2, f_2, 1)$. The relation (3) between the undistorted image points $\mathbf{p}^u_{2i}$ and $\mathbf{p}^u_{1i}$ can be expressed using a 3×3 homography matrix $\mathbf{H}$ as

$$\mathbf{p}^u_{2i} \sim \mathbf{H}\mathbf{p}^u_{1i}, \tag{4}$$

where $\mathbf{H} = \mathbf{K}_2\mathbf{R}_2^\top\mathbf{R}_y\mathbf{R}_1\mathbf{K}_1^{-1}$, and $\sim$ indicates the equality up to a scale factor. The scale, which is equal to $\frac{d_{1i}}{d_{2i}}$ can be eliminated by multiplying (4) with the skew symmetric matrix $[\mathbf{p}^u_{2i}]_\times$, resulting in

$$[\mathbf{p}^u_{2i}]_\times\mathbf{H}\mathbf{p}^u_{1i} = 0. \tag{5}$$

Moreover, equation (5) can be multiplied by $\frac{f_1}{f_2}$ resulting in

$$[\mathbf{p}^u_{2i}]_\times\mathbf{G}\mathbf{p}^u_{1i} = 0, \tag{6}$$

where

$$\mathbf{G} = \widetilde{\mathbf{K}}_2\mathbf{R}_2^\top\mathbf{R}_y\mathbf{R}_1\widetilde{\mathbf{K}}_1^{-1}, \tag{7}$$

with $\widetilde{\mathbf{K}}_2 = \frac{1}{f_2}\mathbf{K}_2 = \mathrm{diag}(1, 1, w_2)$, $w_2 = \frac{1}{f_2}$ and $\widetilde{\mathbf{K}}_1^{-1} = f_1\mathbf{K}_1^{-1} = diag(1, 1, f_1)$. Since $\widetilde{\mathbf{K}}_2$ contains unknowns only in its last row and column, one equation from (6) does not contain $w_2 = \frac{1}{f_2}$ and λ_2.

The rotation matrix $\mathbf{R}_y$ can be parameterized using the Cayley parameterization, which results in a degree-2 polynomial matrix with only one parameter as follows:

$$\mathbf{R}_y = \frac{1}{1+s^2}\begin{bmatrix} 1-s^2 & 0 & 2s \\ 0 & 1+s^2 & 0 \\ -2s & 0 & 1-s^2 \end{bmatrix}. \tag{8}$$

In this case, s corresponds to $\tan\frac{\theta}{2}$, and θ is the yaw angle. Hence, $\cos\theta = \frac{1-s^2}{1+s^2}$ and $\sin\theta = \frac{2s}{1+s^2}$. Since (5) is homogeneous, the scale factor $\frac{1}{1+s^2}$ in (8) can be ignored. Note that (8) introduces a degeneracy for a 180° rotation. However, it was shown in [32, 9, 23] that this degeneracy is not a problem in practice. An alternative formulation is to use the cosines and sines parameterization, which needs two parameters and produces an extra trigonometric identity constraint and, thus, a slower solver with more solutions. Therefore, we only focus on the Cayley parameterization in this paper. Our goal is to estimate the rotation, focal lengths and, potentially, radial distortion parameters of two cameras using the minimal number of point correspondences.

3. Minimal Solutions

In this section, we present new minimal solutions to four practical camera configurations, *i.e.*, solvers H1(G), H1f(G), H1λ(G), H2λf(G). New minimal solvers H2$f_{1,2}$(G) and H3$\lambda_{1,2}f_{1,2}$(G) for varying focal lengths and radial distortions are described in the Supp. material (SM).

3.1. Calibrated Case - H1(G)

A simple case arises when the two cameras have known focal length and known or zero radial distortion, *i.e.*, $\lambda_1 = \lambda_2 = 0$. As such, the new H1(G) solver can be used for images that are captured by a calibrated camera or a camera with known focal length (*e.g.*, from EXIF-tag) and negligible distortion, *i.e.*, narrow field-of-view. In practice, H1(G) can also be used for images captured by wide-angle smartphone cameras, which are often undistorted automatically.

This is a 1-DOF problem with respect to s in $\mathbf{R}_y$ (8). Since (6) contains a skew symmetric matrix of rank 2, each point correspondence gives two linearly independent constraints (6). Therefore, one point correspondence leads to an over-constrained problem. One way of solving this problem is to use a single equation from (6). Such a solver has to find the roots of one quadratic equation in one unknown. However, it can be sensitive to noise. Therefore, we formulate this problem in a least square sense as

$$\min_s (e_1^2(s) + e_2^2(s) + e_3^2(s)), \tag{9}$$

where $e_i(s)$, $i = 1, 2, 3$ are three quadratic polynomials in s from matrix equation (6). The H1(G) solver finds the solution to (9) by computing all stationary points of $e(s) = e_1^2(s) + e_2^2(s) + e_3^2(s)$ and selecting the solution that minimizes (9). This leads to solving one univariate polynomial of degree three, *i.e.*, $\frac{\mathrm{d}e(s)}{\mathrm{d}s} = 0$.

3.2. Common Focal Length Solver - H1f(G)

In the next configuration, we consider two cameras with common unknown focal length, *i.e.*, $f_1 = f_2 = f$, and known or zero radial distortion, *i.e.*, $\lambda_1 = \lambda_2 = 0$. This is a practical scenario that appears in many situations when we have images taken by an uncalibrated camera with fixed unknown focal length and with negligible distortion.

This is a 2-DOF problem with respect to $\{s, f\}$ that can be solved from one point correspondence. Given one point correspondence, constraint (6) leads to three equations from which only two are linearly independent (because of the matrix $[\mathbf{p}_{2i}^u]_\times$). These equations have the following form

$$\boldsymbol{a_1} \cdot [s^2fw, s^2f, s^2w, s^2, sfw, sf, sw, s, fw, f, w, 1]^\top = 0, \tag{10}$$
$$\boldsymbol{a_2} \cdot [s^2fw, s^2f, s^2w, s^2, sfw, sf, sw, s, fw, f, w, 1]^\top = 0, \tag{11}$$
$$\boldsymbol{a_3} \cdot [s^2f, s^2, sf, s, f, 1]^\top = 0, \tag{12}$$

where $\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}$ are coefficient vectors that can be computed from the point correspondence and the gravity direction and $w = 1/f$. Note, that equations (10) and (11) can be multiplied with $f = 1/w$ to obtain polynomial equations in two unknowns $\{s, f\}$ and 9 monomials $\{s^2f^2, s^2f, s^2, sf^2, sf, s, f^2, f, 1\}$. Using (12), f can be expressed as a rational function in s. Substituting this expression into (10) or (11) and multiplying it with the denominator gives us an univariate polynomial in s of degree 6. This polynomial has, however, always the factor $1 + s^2$ that can be eliminated. In this way, we obtain a quartic equation in s, which can be solved in closed-form. Finally, there are up to 4 possible solutions for $\mathbf{R}_y$ and f.

3.3. Common Radial Distortion Solver - H1λ(G)

Another practical case occurs when the two cameras have known focal lengths and an equal but unknown radial distortion, *i.e.*, $\lambda_1 = \lambda_2 = \lambda$. This happens when the images are captured by a fixed wide-angle camera whose focal length can be extracted from the EXIF-tag.

This is a 2-DOF problem with respect to $\{s, \lambda\}$. Similar to Sec. 3.2, one point correspondence in (6) gives three equations (two linearly independent ones) of the form

$$\boldsymbol{a_1} \cdot [s^2\lambda^2, s^2\lambda, s^2, s\lambda^2, s\lambda, s, \lambda^2, \lambda, 1]^\top = 0, \tag{13}$$
$$\boldsymbol{a_2} \cdot [s^2\lambda^2, s^2\lambda, s^2, s\lambda^2, s\lambda, s, \lambda^2, \lambda, 1]^\top = 0, \tag{14}$$
$$\boldsymbol{a_3} \cdot [s^2\lambda, s^2, s\lambda, s, \lambda, 1]^\top = 0, \tag{15}$$

where $\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}$ are coefficient vectors that can be computed from the point correspondence and the gravity direction. Using (15), λ can be expressed as a rational function in s. Substituting this expression into (13) or (14) and multiplying it with the denominator gives us an univariate polynomial in s of degree 6. This polynomial has again the factor $1 + s^2$ that can be eliminated, resulting in a quartic equation in s. Finally, there are up to 4 possible solutions.

3.4. Common Focal Length, Distortion - H2λf(G)

Finally, we address the problem where the images are taken by a camera with fixed unknown focal length and radial distortion, *i.e.* $f_1 = f_2 = f$ and $\lambda_1 = \lambda_2 = \lambda$ in (6). In this case, there are three unknowns $\{s, f, \lambda\}$, and we need at least 1.5 point correspondences. In practice, we still need

	H4	H4f(G)	H2	H2f	H5λ	H3λf	H1(G)	H1f(G)	H1λ(G)	H2λf(G)
Reference	[15]	[8]	[16]	[5]	[10]	[19][6]				
Radial distortion					✓	✓			✓	✓
No. of points	4	4	2	2	5	3	1	1	1	2
No. of solutions	1	24	1	3	18	18	1(1)	4(2)	4(2)	6(2)
Gravity prior		✓					✓	✓	✓	✓
Pure $\mathbf{R}$			✓	✓		✓	✓	✓	✓	✓
DOF	8	7	3	4	9	5	1	2	2	3

Table 1: The properties of the proposed gravity-based (gray) and state-of-the-art solvers. The number of solutions in brackets refers to a special case when the y-axis of the camera is considered to be physically aligned with the gravity direction.

to sample 2 points, but we only need 3 out of the 4 linearly independent equations. The 4^{th} equation can be used to eliminate geometrically infeasible solutions.

Two out of the three equations from (6) are of degree 6 and one is of degree 4, *i.e.*, the equation of degree four corresponds to the last row of matrix $[\mathbf{p}_{2i}^{u}]_{\times}$ that does not contain λ and multiplies only the first two rows of $\widetilde{\mathbf{K}}_2$, which do not contain f. Therefore, to solve the problem we use equations of degree 6 and 4 from the first correspondence and the equation of degree 4 from the second correspondence. We obtain the following three equations

$$\begin{aligned} \boldsymbol{a_1} \cdot [&s^2f^2\lambda^2, s^2f^2\lambda, s^2f^2, s^2f\lambda, s^2f, s^2, sf^2\lambda^2, \\ &sf^2\lambda, sf^2, sf\lambda, sf, s, f^2\lambda^2, f^2\lambda, f^2, f\lambda, f, 1] = 0, \end{aligned} \tag{16}$$

$$\boldsymbol{a_2} \cdot [s^2f\lambda, s^2f, s^2, sf\lambda, sf, s, f\lambda, f, 1] = 0, \tag{17}$$

$$\boldsymbol{a_3} \cdot [s^2f\lambda, s^2f, s^2, sf\lambda, sf, s, f\lambda, f, 1] = 0. \tag{18}$$

Parameter λ always appears together with f, so we let $\tau = f\lambda$. In this case, the three equations above are written as

$$\begin{aligned} \boldsymbol{a_1} \cdot [&s^2\tau^2, s^2f\tau, s^2f^2, s^2\tau, s^2f, s^2, s\tau^2, \\ &sf\tau, sf^2, s\tau, sf, s, \tau, f\tau, f^2, \tau, f, 1] = 0, \end{aligned} \tag{19}$$

$$\boldsymbol{a_2} \cdot [s^2\tau, s^2f, s^2, s\tau, sf, s, \tau, f, 1] = 0, \tag{20}$$

$$\boldsymbol{a_3} \cdot [s^2\tau, s^2f, s^2, s\tau, sf, s, \tau, f, 1] = 0. \tag{21}$$

This system in unknowns $\{s, f, \tau\}$ can be solved using the Gröbner basis method. Using the automatic generator [23], we obtained a solver that performs G-J elimination of a template matrix of size 30×38, and eigendecomposition of an 8×8 action matrix, *i.e.*, the system has 8 solutions. There are always two infeasible imaginary solutions of the form $s^2 = -1$ and only up to 6 possible real ones.

An alternative and more efficient way of solving such a system is to eliminate f, τ from the original equations. Using (20) and (21), τ and f can be expressed as rational functions in s. Then substituting the formulations of τ and f into (19), we obtain a univariate polynomial in s of degree 8. This polynomial has again the factor $1 + s^2$ that can be eliminated, resulting in sextic equation in s, which can be efficiently solved using Sturm sequences [13].

3.5. Special case

If the y-axis of the camera is considered to be physically aligned with the gravity direction, *i.e.*, $\mathbf{R}_1, \mathbf{R}_2$ are identity matrices, there are many zero coefficients in the systems in our solvers. In such a scenario, all four solvers are very simple and result in one quadratic equation. The properties of all the stitching solvers are shown in Table 1.

4. Experiments

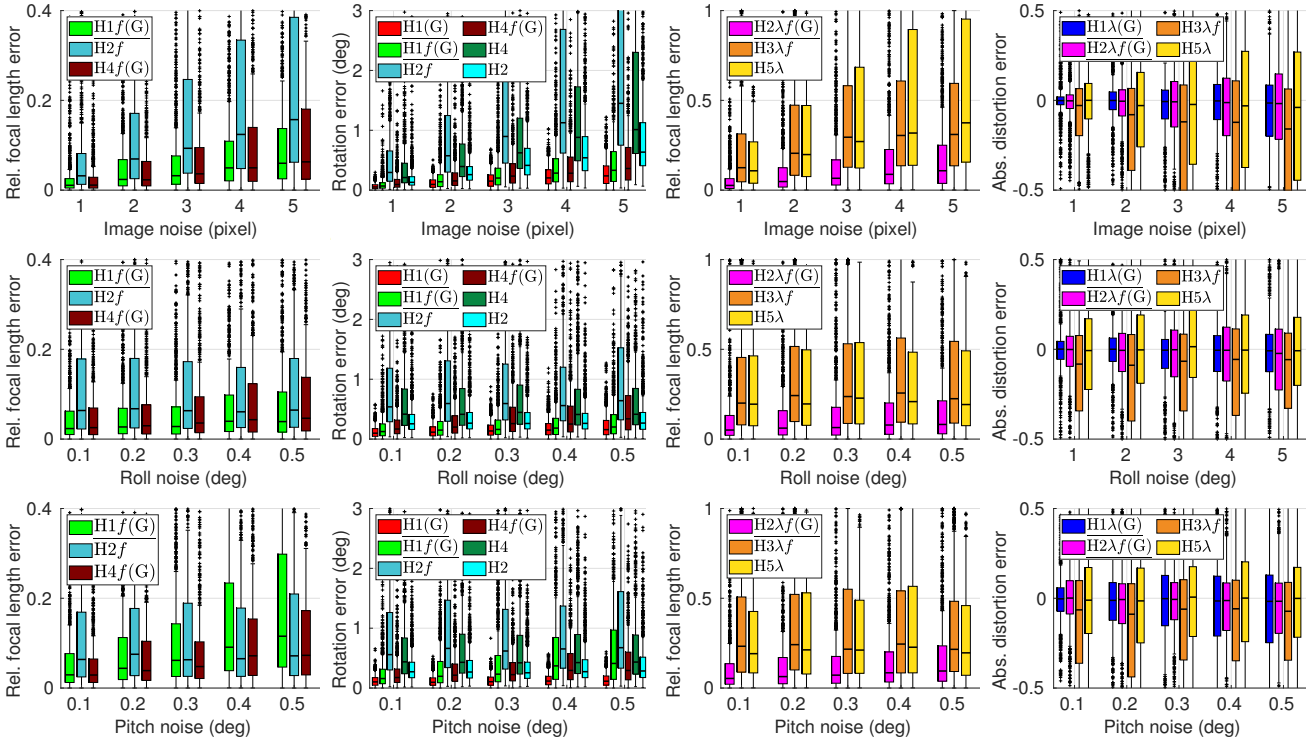
In this section, we study the performance of the proposed H1(G), H1f (G), H1λ (G), and H2λf(G) solvers on both synthetic and real images. For comparison, the following state-art-solvers are considered: **(1)** general 4pt homography solver H4 [15]; **(2)** homography with unknown f and known gravity H4f(G) [8]; **(3)** calibrated stitching solver H2 [16]; **(4)** stitching solver with unknown f, H2f [5]; **(5)** homography with unknown distortion λ, H5λ [10]; **(6)** stitching with unknown f and λ, H3λf [19, 6]. Since [6] solves the same problem as [19], but the method in [6] is more stable than [19], we included [6] in our experiments. All solvers, together with their properties and the color coding used in the experiments, are summarized in Table 1.

Experiments with our solvers H2$f_{1,2}$(G) and H3$\lambda_{1,2}f_{1,2}$(G) for varying focal lengths and radial distortions and state-of-the-art solvers H3$f_{1,2}$ [5], H5$\lambda_{1,2}$ [22] and H6$\lambda_{1,2}$ [22] are in the SM. All solvers get randomly selected minimal samples of their respective sizes as input.

4.1. Synthetic Evaluation

We choose the following setup to generate synthetic data. 200 points are randomly distributed in the box $[-3, 3] \times [-3, 3] \times [4, 6]$ in the first camera's local coordinate system. A random but feasible rotation is applied to the points. The focal length f_g is set to 1000 pixels. We generate 1000 pairs of images with different rotations. The focal length error is $|f_e - f_g| / f_g$, where f_e is the estimated focal length. The rotation error is $\epsilon_R = \arccos\left(\left(\text{tr}\left(\mathbf{R_e}\mathbf{R_g}^\top\right) - 1\right)/2\right)$, where $\mathbf{R_e}$ is the estimated and $\mathbf{R_g}$ is the ground truth 3-DoF rotation matrix. This is a fair comparison since it accounts for the error in the gravity direction as well, that is incorporated in $\mathbf{R_e}$ when the proposed solvers are used.

Fig. 3a reports the focal length (left column) and rotation errors (right) of the solvers assuming zero distortion. The top row shows the performance under increasing im-



(a) No radial distortion.

(b) Radial distortion.

Figure 3: **(a)** Boxplot of the relative focal length error and rotation error for the zero distortion case. **Left column**: Relative focal length error. **Right column**: Rotation error. **(b)** Boxplot of the relative focal length error and absolute distortion error for the distortion solvers. **Left column**: Relative focal length error. **Right column**: Absolute distortion error. From top to bottom: increased image measurement noise, increased roll noise with constant 2 pixel standard deviation image measurement noise, increased pitch noise with constant 2 pixel standard deviation image measurement noise.

age noise with different standard deviations. The proposed H1(G) (red) and H1f (green) solvers perform significantly better than the other ones under varying image noise.

Since, in real applications, the alignment of the camera or the gravity vector measured by the IMU is not perfect, we also added noise to the roll and pitch angles. The middle and bottom rows show the performance with increasing roll and pitch noise under constant image noise of 2 pixel standard deviation. Our solvers are comparable to the state-of-the-art methods even with roll and pitch noise up to 0.5°. The upper bound of the noise is chosen to follow the noise from the lower grade MEMS IMUs [24]. Nowadays, accelerometers used in modern smartphones and camera-IMU systems have noise levels around 0.06° and expensive "good" ones have lower than 0.02° [12].

We also study the performance of the radial distortion solvers. The distortion parameter was set to $\lambda_g = -0.4$, corresponding to medium distortion. The error was defined as $\lambda_g - \lambda_e$, where λ_e is the estimated distortion. Fig. 3b reports the relative focal length (left) and absolute radial distortion error (right). Our new solvers H1λ(G) (blue) and H1λf(G) (magenta) outperform the state-of-the art ones H3λf and H5λ that are very sensitive to noise. For H5λ, we extracted the focal length using the method from [5].

4.2. Real-world Experiments

To test the proposed techniques on real-world data, we chose the SUN360 [34] panorama dataset. The purpose of the SUN360 database is to provide academic researchers a comprehensive collection of annotated panoramas covering 360×180-degree full view for a large variety of environmental scenes, places and the objects within. To build the core of the dataset, high-resolution panorama images were downloaded and grouped into different place categories.

To obtain radially distorted image pairs from each 360° panoramic scene, we cut out images simulating a 80° FOV camera with a step size of 10°. Thus, the rotation around the vertical axis between two consecutive images is always 10°. Finally, image pairs were formed by pairing all images with a common field-of-view in each scene. In total, $579,800$ image pairs were generated. Moreover, to test the methods also in cases when there is no distortion, we undistorted the images using the ground truth distortion parameters. In each image, 8000 SIFT keypoints are detected in order to have a reasonably dense point cloud and a stable result [20]. We combined a mutual nearest neighbor check with the stan-

dard distance ratio test [26] to establish tentative correspondences [20]. To get ground truth correspondences which can be used to calculate the re-projection error, we composed the ground truth homography and selected its inliers given the used inlier-outlier threshold. Methods not estimating the focal length were provided the ground truth.

To test the solvers on real-world data, we chose GC-RANSAC [4]. In GC-RANSAC, two different solvers are used: (a) one for estimating the homography from a minimal sample and (b) one for fitting to a larger-than-minimal sample when doing final homography polishing on all inliers or in the local optimization step. For (a), the objective is to solve the problem using as few correspondences as possible since the processing time depends *exponentially* on the number of correspondences required for the estimation. The proposed solvers are included in this step. For (b), when testing the methods on the distorted images, we applied the H6$\lambda_{1,2}$ solver [22] to estimate the homography and distortion parameters from a larger-than-minimal sample. Therefore, even minimal solvers not estimating the radial distortion obtain reasonably accurate solutions. In case when the undistorted images are used, we applied the normalized H4 [15] homography estimator. The inlier threshold was set to 3 pixels. Note that the situation where the y-axes of the cameras are physically aligned with the gravity direction, *i.e.*, $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}$ with pure rotation, is a degenerate configuration for the H4f(G) [8] solver. In such cases this solver does not give meaningful results. Since our real data was close to this degenerate situation (approximately aligned) we exclude H4f(G) in the real experiments.

Table 2 reports the average processing times (in ms), re-projection error (in pixels) and focal length error (in %) both on the distorted and undistorted images. The corresponding cumulative distribution functions (CDF) on the distorted images are shown in Fig. 4. Being accurate is interpreted as a curve close to the top-left corner. CDFs for undistorted images are in the supplementary material.

Undistorted images. After the images have been undistorted, all methods lead to fairly similar re-projection errors. In the calibrated case, the proposed H1(G) solver is the fastest while leading to similarly accurate results as the other solvers. Among the methods that estimate the focal length, the proposed H1f(G) solver leads to the most accurate focal lengths while being the fastest.

Distorted images. When having images with unknown radial distortion and known focal length, the proposed H1λ(G) solvers obtains the most accurate stitching results in terms of re-projection error while, simultaneously, being the fastest. In the unknown focal length case, the proposed H1f(G) solver has the lowest run-time and it is the second most accurate algorithm. In this case, the H3λf method is the most accurate while, however, it is also the slowest.

Smartphone images. To further show the benefits of the proposed solvers, we tested them on the data recorded from two devices (iPhone 6s, iPhone 11 pro) with three cameras: the wide-angle cameras of the iPhone 6s and iPhone 11 with focal lengths of 29mm and 26mm, respectively, and the ultra wide-angle camera of the iPhone 11 with a focal length of 13mm. The sequences were captured at 1280x720@30Hz with the IMU readings captured at 100Hz. The images and IMU data were synchronized based on their timestamps. 4240 image pairs for the two wide-angle cameras and 3530 image pairs for the ultra wide-angle camera with synchronized gravity direction were obtained. The ground truth focal lengths were obtained by offline calibration of the cameras. Since we do not know the ground truth homography and, thus, ground truth inliers, we do not measure the re-projection error on these datasets. The CDFs of the focal length errors and processing times are shown in Fig. 5a. On these experiments, the proposed H1f(G) solver leads to the most accurate focal lengths while being the fastest method as well. The other one-point solvers are also very efficient compared to the other tested algorithms.

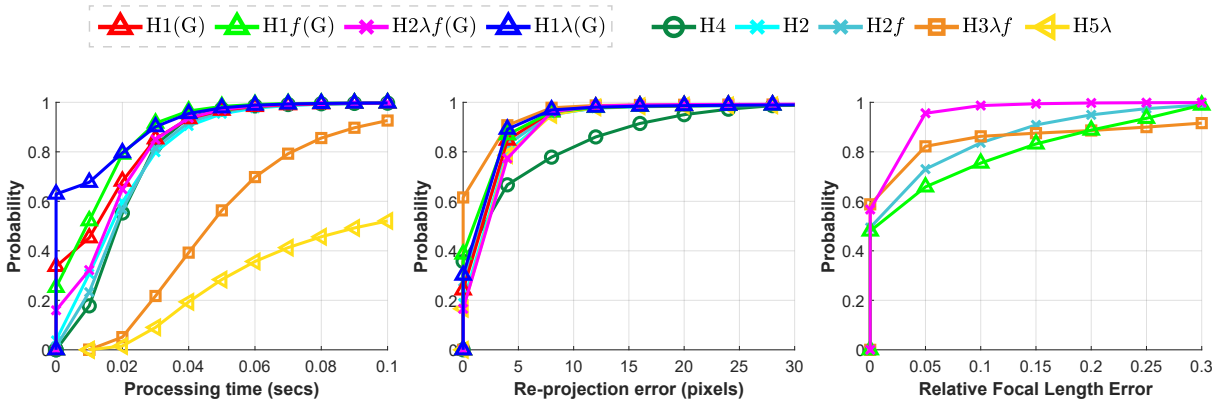
	Undistorted SUN360 images			Distorted SUN360 images		
	ϵ_r (px)	ϵ_f (%)	t (ms)	ϵ_r (px)	ϵ_f (%)	t (ms)
H1(G)	**0.7**	–	**9.1**	3.0	–	**19.3**
H1λ(G)	**0.7**	–	11.6	**2.6**	–	**12.2**
H2	**0.7**	–	**10.9**	3.5	–	25.1
H4	**0.7**	–	13.0	**2.9**	–	26.5
H5λ	**0.7**	–	104.1	3.7	–	186.1
H1f(G)	**0.7**	**0.1**	**7.8**	**2.4**	2.9	**16.5**
H2f	**0.7**	**0.1**	**12.4**	3.1	2.6	25.9
H2λf(G)	**0.7**	**0.3**	15.0	3.8	**1.7**	**22.3**
H3λf	**0.6**	0.4	46.6	**1.1**	**1.0**	59.1

Table 2: The avg. run-time (t; ms), median re-projection (ϵ_r; px) and focal length (ϵ_f; %) errors on 579,800 image pairs from the SUN360 dataset. Top part: known focal length solvers. Bottom part: unknown focal length solvers. Best and second best results are shown, respectively, in red and blue. The corresponding CDFs are in Fig 4.

4.3. Computational Complexity

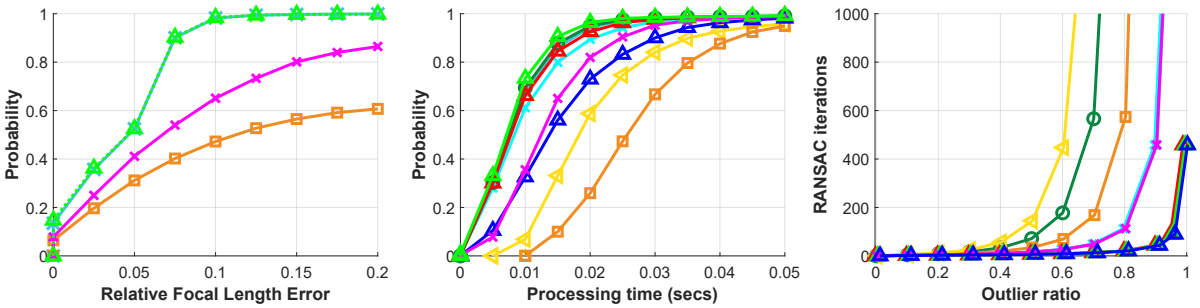
The complexity and run-time of a single estimation of the compared solvers are reported in the following table.

Solver	G-J	Eigen	Poly	Time (μs)
H4	8×9	-	-	8
H4f(G)	14×33	33×33	-	121
H2	-	-	-	6
H2f	-	3×3	-	6
H5λ	9×18	18×18	-	40
H3λf	90×132	25×25	-	80
H1(G)	-	-	3	4
H1f(G)	-	-	4	5
H1λ(G)	-	-	4	5
H2λf(G)	-	-	6	5



(a) Distorted images

Figure 4: The cumulative distribution functions of the processing times (in seconds), average re-projection errors (in pixels) and relative focal length errors of GC-RANSAC [4] when combined with different minimal solvers. The values are calculated from a total of $579,800$ image pairs from the SUN360 dataset. The confidence was set to 0.99 and the inlier threshold to 3 px. Being accurate is interpreted as a curve close to the top-left corner. The corresponding values are reported in Table 2.



(a) The captured smartphone dataset.

(b) Theoretical RANSAC iterations.

Figure 5: **(a)** The cumulative distribution functions of the processing time (in seconds) and relative focal length errors of GC-RANSAC [4] when combined with different minimal solvers. The values are calculated from a total of $7,770$ image pairs from the captured phone dataset. Being accurate is interpreted as a curve close to the top-left corner. **(b)** The theoretical number of RANSAC iterations for each solver plotted as the function of the outlier ratio. The confidence was set to 0.99.

We only show the major steps performed by each solver. The number in the cells, *e.g.* 9×18, denotes the matrix size to which the G-J elimination or eigendecomposition is applied. The number in the fourth column denotes the degree of the univariate polynomial that needs to be solved. A table for the varying $f_{1,2}$ and $\lambda_{1,2}$ solvers is in the SM.

The theoretical number of RANSAC iterations is shown in Fig 5b plotted as the function of the outlier ratio in the data. It can be seen that the proposed solvers, due to requiring the fewest correspondences, lead to a small number of RANSAC iterations even in the most challenging cases. Note that we added a small random offset on axis x to make sure the curves do not overlap entirely.

5. Conclusions

In this paper, we propose six new minimal solvers for image stitching, considering different configurations of cameras with coinciding optical axes and aligned with the gravity direction. These configurations include solvers for a fully calibrated camera, a camera with unknown fixed or varying focal length and with or without radial distortion. The proposed methods are tested on synthetic scenes and on more than 500k image pairs generated from the spherical images of the SUN360 dataset both on radially distorted and undistorted images. Since we have not found datasets for image stitching with available gravity vector, we captured a new dataset with two different smartphones (three different cameras) equipped with IMU sensors. The dataset consists of 7770 image pairs in total. We show that the new solvers have similar or better accuracy than the state-of-the-art solvers and outperform them in terms of processing time. The source code and dataset are available at https://github.com/yaqding/stitching-gravity

Acknowledgments. This research was supported by the National Science Fund of China (Grant No. U1713208), the "111" Program B13022, and the ERC-CZ grant MSMT LL1901.

References

[1] Lourdes Agapito, Eric Hayman, and Ian Reid. Self-calibration of rotating and zooming cameras. *International journal of computer vision*, 45(2):107–127, 2001. 3

[2] Cenek Albl, Zuzana Kukelova, and Tomas Pajdla. Rolling shutter absolute pose problem with known vertical direction. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[3] Clemens Arth, Manfred Klopschitz, Gerhard Reitmayr, and Dieter Schmalstieg. Real-time self-localization from panoramic images on mobile devices. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 37–46, 2011. 1

[4] Daniel Barath and Jiří Matas. Graph-cut RANSAC. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 7, 8

[5] Matthew Brown, Richard I Hartley, and David Nistér. Minimal solutions for panoramic stitching. In *Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 2, 5, 6

[6] Martin Byröd, Matthew A Brown, and Kalle Åström. Minimal solutions for panoramic stitching with radial distortion. In *British Machine Vision Conference (BMVC)*, 2009. 1, 2, 3, 5

[7] Jianhui Chen, Fangrui Zhu, and James J Little. A two-point method for PTZ camera calibration in sports. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 287–295. IEEE, 2018. 1, 3

[8] Yaqing Ding, Jian Yang, Jean Ponce, and Hui Kong. An efficient solution to the homography-based relative pose problem with a common reference direction. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 5, 7

[9] Yaqing Ding, Jian Yang, Jean Ponce, and Hui Kong. Minimal solutions to relative pose estimation from two views sharing a common direction with unknown focal length. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4

[10] Andrew Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Computer Vision and Pattern Recognition (CVPR)*, 2001. 1, 2, 3, 5

[11] Jan-Michael Frahm and Reinhard Koch. Camera calibration with known rotation. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1418–1418. IEEE Computer Society, 2003. 2

[12] Friedrich Fraundorfer, Petri Tanskanen, and Marc Pollefeys. A minimal case solution to the calibrated relative pose problem for the case of two known orientation angles. In *European Conference on Computer Vision (ECCV)*, 2010. 2, 6

[13] Walter Gellert, M Hellwich, H Kästner, and H Küstner. *The VNR concise encyclopedia of mathematics*. Springer Science & Business Media, 2012. 5

[14] Richard Hartley and Hongdong Li. An efficient hidden variable approach to minimal-case camera motion estimation. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2012. 3

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2, 5, 7

[16] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Josa a*, 1987. 2, 5

[17] Manish Jethwa, Andrew Zisserman, and Andrew W Fitzgibbon. Real-time panoramic mosaics and augmented reality. In *BMVC*, pages 1–11, 1998. 1

[18] Shunping Ji, Zijie Qin, Jie Shan, and Meng Lu. Panoramic slam from a multiple fisheye camera rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:169–183, 2020. 1

[19] Hailin Jin. A three-point minimal solution for panoramic stitching with lens distortion. In *Computer Vision and Pattern Recognition (CVPR)*, 2008. 1, 2, 3, 5

[20] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision (IJCV)*, 2020. 6, 7

[21] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-form solutions to minimal absolute pose problems with known vertical direction. In *Asian Conference on Computer Vision (ACCV)*, 2010. 2

[22] Zuzana Kukelova, Jan Heller, Martin Bujnak, and Tomas Pajdla. Radial distortion homography. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 5, 7

[23] Viktor Larsson, Kalle Åström, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 5

[24] Gim Hee Lee, Marc Pollefeys, and Friedrich Fraundorfer. Relative pose estimation for a multi-camera system with known vertical direction. In *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 6

[25] Thomas Lemaire and Simon Lacroix. Slam with panoramic vision. *Journal of Field Robotics*, 24(1-2):91–111, 2007. 1

[26] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, 1999. 7

[27] Andrew MacQuarrie and Anthony Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*, pages 45–54. IEEE, 2017. 1

[28] Oleg Naroditsky, Xun S Zhou, Jean Gallier, Stergios I Roumeliotis, and Kostas Daniilidis. Two efficient solutions for visual odometry using directional correspondence. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2012. 2

[29] Marcus Valtonen Ornhag, Patrik Persson, Marten Wadenback, Kalle Astrom, and Anders Heyden. Efficient real-time radial distortion correction for uavs. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1751–1760, 2021. 2

[30] Olivier Saurer, Pascal Vasseur, Rémi Boutteau, Cédric Demonceaux, Marc Pollefeys, and Friedrich Fraundorfer. Homography based egomotion estimation with a common direction. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2017. 2

[31] Chris Sweeney, John Flynn, Benjamin Nuernberger, and Matthew Turk. Efficient computation of absolute pose for gravity-aware augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality*, 2015. 2

[32] Chris Sweeney, John Flynn, and Matthew Turk. Solving for relative pose with a partially known rotation is a quadratic eigenvalue problem. *International Conference on 3D Vision (3DV)*, 2014. 2, 4

[33] Lang Wang, Wen Yu, and Bao Li. Multi-scenes image stitching based on autonomous driving. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020. 1

[34] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[35] Tao Yang, Zhi Li, Fangbing Zhang, Bolin Xie, Jing Li, and Linfeng Liu. Panoramic UAV surveillance and recycling system based on structure-free camera array. *IEEE Access*, 7:25763–25778, 2019. 1