

A Fully Projective Formulation to Improve the Accuracy of Lowe’s Pose–Estimation Algorithm ^{*}

Helder Araújo

Rodrigo L. Carceroni

Christopher M. Brown

University of Rochester
Computer Science Department
Rochester, NY – 14627 – USA

Abstract

Both the original version of David Lowe’s influential and classic algorithm for tracking known objects and a reformulation of it implemented by Ishii *et al.* rely on (different) approximated imaging models. Removing their simplifying assumptions yields a fully projective solution with significantly improved accuracy and convergence, and arguably better computation–time properties.

1 Introduction and History

The ability to track a set of points in a moving image plays a fundamental role in several computer vision applications with real–time constraints such as autonomous navigation, surveillance, grasping, manipulation and augmented reality. Often some geometrical invariants of these points (such as their relative spatial positions, in the case of a rigid object) are known in advance. Algebraic solutions with perspective camera models have been proposed for several variations of this problem [1; 22; 17; 8; 20; 5; 10; 12; 6]. However, the resulting techniques usually work only with a limited number of points and are thus sensitive to additive noise and erroneous matching. Furthermore, they usually depend on numerical techniques for finding zeros of fourth–or–higher–degree polynomial equations.

^{*}This material is based on work supported by the Luso–American Foundation, Calouste Gulbenkian Foundation, JNICT, CAPES process BEX 0591/95-5, NSF IIP grant CDA-94-01142, NSF grant IRI-9306454 and DARPA grant DAAB07-97-C-J027.

Pioneering work by Lowe [15; 14; 13] and Gennery [7] addressed the problem in a projective framework. Lowe showed that the direct use of numerical optimization techniques is an effective way to overcome the lack of robustness that makes the traditional analytical techniques infeasible in practice.

DeMenthon and Davis [4; 18] and Horaud *et al.* [9] propose techniques that start with weak– or para–perspective solutions, respectively, and refine them iteratively to recover the full–perspective pose. Phong, Horaud *et al.* [19] showed that it is possible to decouple completely the recovery of rotational pose parameters from their translational counterparts. However, unlike Lowe’s, none of these methods is easily generalizable to deal with uncalibrated focal length or objects (scenes) with internal degrees of freedom.

Lowe’s algorithm is attractive because of its elegant simplicity and its powerful generality. In this note, we first recall the original algorithm and another incarnation from the literature. Both algorithms contain certain simplifying assumptions that are easily eliminated. We present and comparatively evaluate the resulting fully projective solution. It preserves the appealing properties of Lowe’s original conception while performing substantially better than either approximation. Section 7 relates our findings to previous speculations on and analyses of Lowe’s algorithm. This note is an abbreviation of [2], which is less terse and contains more experimental results.

2 Lowe’s Algorithm

Lowe’s original algorithm [15; 14; 13] addresses the issue of viewpoint and model parameter computation,

given a known 3-D object and the corresponding image. It assumes that the imaging process is a projective transformation. The method can thus be used to identify the pose (translation and orientation with respect to the camera coordinate system) of a local coordinate system affixed to an imaged rigid object. It can also be extended to discover the values of other parameters such as the camera focal length and shape parameters of non-rigid objects. The recovery process is based on the application of Newton’s method.

Rather than solving directly for the parameter vector \mathbf{s} in a nonlinear system, Newton’s method computes a vector of corrections δ to be subtracted from the current estimate for \mathbf{s} on each iteration. If $\mathbf{s}^{(i)}$ is the parameter vector for iteration i , then:

$$\mathbf{s}^{(i+1)} = \mathbf{s}^{(i)} - \delta. \quad (1)$$

Given a vector of error measurements \mathbf{e} between components of the model and the image, we want to solve for a correction vector δ that eliminates this error:

$$\mathbf{J} \delta = \mathbf{e}, \quad \text{where: } \mathbf{J}_{ij} = \frac{\partial e_i}{\partial x_j}. \quad (2)$$

The equations used to describe the projection of a three-dimensional model point \mathbf{p} into a two-dimensional image point $[u, v]$ are:

$$\begin{aligned} [x, y, z]^T &= \mathbf{R}(\mathbf{p} - \mathbf{t}), \\ [u, v] &= f \begin{bmatrix} \frac{x}{z} & \frac{y}{z} \end{bmatrix}, \end{aligned} \quad (3)$$

where T denotes transpose, \mathbf{t} is a 3-D translation vector (defined in the model coordinate frame) and \mathbf{R} is a rotation matrix that transforms \mathbf{p} in the original model coordinates into a point $[x, y, z]^T$ in camera-centered coordinates. These are combined in the second equation above with the focal length f to perform perspective projection into an image point $[u, v]$.

The problem is to solve for \mathbf{t} , \mathbf{R} and possibly f , given a number of model points and their corresponding locations in an image. In order to apply Newton’s method, we must be able to calculate the partial derivatives of u and v with respect to each of the unknown parameters. Lowe [14] proposes a reparameterization of the projection equations, to simplify the calculation by “express[ing] the translations in terms of the camera coordinate system rather than model

coordinates”:

$$\begin{aligned} [x', y', z']^T &= \mathbf{R} \mathbf{p}, \\ [u, v] &= \left[f \frac{x'}{z' + d_z} + d_x, f \frac{y'}{z' + d_z} + d_y \right]. \end{aligned} \quad (4)$$

The variables \mathbf{R} and f remain the same as in the previous transform, but vector \mathbf{t} has been replaced by the parameters d_x , d_y and d_z . The two transforms are equivalent when:

$$\mathbf{t} = -\mathbf{R}^{-1} \left[\frac{d_x(z' + d_z)}{f}, \frac{d_y(z' + d_z)}{f}, d_z \right]^T. \quad (5)$$

According to Lowe, “in the new parameterization, d_x and d_y simply specify the location of the object on the image plane and d_z specifies the distance of the object from the camera”. To compute the partial derivatives of the error with respect to the rotation angles (ϕ_x , ϕ_y and ϕ_z are the rotation angles about x , y and z , respectively), it is necessary to calculate the partial derivatives of x , y and z with respect to these angles. Table 1 gives these derivatives for all combinations of variables.

	x	y	z
ϕ_x	0	$-z'$	y'
ϕ_y	z'	0	$-x'$
ϕ_z	$-y'$	x'	0

Table 1: The partial derivatives of x , y and z with respect to counterclockwise rotations ϕ (in radians) about the coordinate axes.

Newton’s method is carried out by calculating the optimum correction rotations $\Delta\phi_x$, $\Delta\phi_y$ and $\Delta\phi_z$ to be made about the camera-centered axes. Given Lowe’s parameterization, the partial derivatives of u and v with respect to each of the seven parameters of the imaging model (including the focal length f) are given by Table 2.

Lowe then notes that each iteration of the multi-dimensional Newton’s method solves for a vector of corrections

$$\delta = [\Delta d_x, \Delta d_y, \Delta d_z, \Delta\phi_x, \Delta\phi_y, \Delta\phi_z]^T. \quad (6)$$

Lowe’s algorithm dictates that for each point in the model matched against some corresponding point

	u	v
d_x	1	0
d_y	0	1
d_z	$-fc^2x'$	$-fc^2y'$
ϕ_x	$-fc^2x'y'$	$-fc(z' + cy'^2)$
ϕ_y	$fc(z' + cx'^2)$	$fc^2x'y'$
ϕ_z	$-fcy'$	fcx'
f	cx'	cy'

Table 2: The partial derivatives of u and v with respect to each of the camera viewpoint parameters and the focal length, according to Lowe’s original approximation. Here $c = \frac{1}{z'+d_z}$.

in the image, we first project the model point into the image using the current parameter estimates and then measure the error in the resulting position with respect to the given image point. The u and v components of the error can be used independently to create separate linearized constraints. Making use of the u component of the error, E_u , we create an equation that expresses this error as the sum of the products of its partial derivatives times the unknown error-correcting values:

$$\frac{\partial u}{\partial d_x} \Delta d_x + \frac{\partial u}{\partial d_y} \Delta d_y + \frac{\partial u}{\partial d_z} \Delta d_z + \frac{\partial u}{\partial \phi_x} \Delta \phi_x + \frac{\partial u}{\partial \phi_y} \Delta \phi_y + \frac{\partial u}{\partial \phi_z} \Delta \phi_z = E_u. \quad (7)$$

The same point yields a similar equation for its v component. Thus each point correspondence yields two equations. As Lowe says: “from three point correspondences we can derive six equations and produce a complete linear system which can be solved for all six camera-model corrections”.

3 Lowe’s Approximation

Lowe’s formulation assumes that d_x and d_y are constants to be determined by the iterative procedure, when in fact they are not constants at all — they depend on the location of the points being imaged.

Let the rows of the rotation matrix \mathbf{R} be denoted

by \mathbf{r}_x , \mathbf{r}_y and \mathbf{r}_z , such that:

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_x \\ \mathbf{r}_y \\ \mathbf{r}_z \end{bmatrix}.$$

Then using the projective transformation formulated in Eq. (3) the new parameters d_x , d_y , d_z are given by:

$$d_z = -\mathbf{r}_z \cdot \mathbf{t}, \text{ and then:} \\ [d_x, d_y] = -f \left[\frac{\mathbf{r}_x \cdot \mathbf{t}}{\mathbf{r}_z \cdot \mathbf{p} + d_z}, \frac{\mathbf{r}_y \cdot \mathbf{t}}{\mathbf{r}_z \cdot \mathbf{p} + d_z} \right]. \quad (8)$$

Notice that d_z is dependent only on the object pose parameters, but d_x and d_y are also a function of each point’s coordinates in the object coordinate frame. It is therefore in general impossible to find a single consistent value either for d_x or for d_y . In the general case both these parameters will depend on the position of each individual object feature. They are not constants — they are only the same for those points for which $\mathbf{r}_z \cdot \mathbf{p}$ has the same value. Therefore we can not use d_x and d_y as defined in Eq. (4). The assumption that is implicit in Lowe’s algorithm as published is that the corrections needed for the translation are much larger than those due to the rotation of the object. However, if no restrictions are imposed, the coordinates of the points in the object coordinate frame (\mathbf{p}) can assume high values. Even if they do not, the term $\mathbf{r}_z \cdot \mathbf{p}$ may change significantly (due to the object’s own geometry) and affect the estimation process.

4 Ishii’s Approximation

Ishii’s formulation [11] also contains simplifications. Image formation is again given by Eq. (3).

Defining:

$$[x_t, y_t, z_t]^T = \mathbf{R} \mathbf{t}, \quad (9)$$

the partial derivatives of u and v with respect to each of the seven parameters of the camera model are given by Table 3. The vector $[x_t, y_t, z_t]^T$ represents the translation vector in the camera coordinate frame. In this approximation the computation of the partial derivatives is performed using the coordinates of the points in the object coordinate frame, **ignoring the effect of rotation**.

	u	v
x_t	$-fc$	0
y_t	0	$-fc$
z_t	fac^2	$fb c^2$
ϕ_x	$-fa c^2 p_y$	$-fc(p_z + bc p_y)$
ϕ_y	$fc(p_z + ac p_x)$	$fb c^2 p_x$
ϕ_z	$-fc p_y$	$fc p_x$
f	ac	bc

Table 3: The partial derivatives of u and v with respect to each of the camera viewpoint parameters and the focal length according to Ishii’s approximation. Here $[a, b, c] = [p_x - x_t, p_y - y_t, \frac{1}{p_z - z_t}]$, where $\mathbf{p} = [p_x, p_y, p_z]^T$.

5 Our Fully Projective Solution

Initially define x' , y' and z' as in Lowe’s formulation:

$$[x', y', z']^T = \mathbf{R}\mathbf{p}.$$

Model the image formation process by Eq. (3). Remove the approximations of Lowe and Ishii by defining:

$$[d'_x, d'_y, d'_z] = -[\mathbf{r}_x \cdot \mathbf{t}, \mathbf{r}_y \cdot \mathbf{t}, \mathbf{r}_z \cdot \mathbf{t}]. \quad (10)$$

In this case the image coordinates of each point are given by:

$$[u, v] = f \left[\frac{x' + d'_x}{z' + d'_z}, \frac{y' + d'_y}{z' + d'_z} \right]. \quad (11)$$

The partial derivatives of u and v with respect to each of the six pose parameters and the focal length are given by Table 4.

As in Lowe’s formulation, the translation vector is computed using Eq. (5), with d'_x , d'_y and d'_z as defined in Eq. (10). This translation vector is defined in the object coordinate frame. The minimization process yields estimates of d'_x , d'_y and d'_z , which are the result of the product of the rotation matrix by the translation vector.

A numerically equivalent but conceptually more elegant way of looking at this solution is through a redefinition of the image formation process, so that rotation and translation are explicitly decoupled, and

	u	v
d'_x	fc	0
d'_y	0	fc
d'_z	$-fac^2$	$-fb c^2$
ϕ_x	$-fa c^2 y'$	$-fc(z' + bc y')$
ϕ_y	$fc(z' + ac x')$	$fb c^2 x'$
ϕ_z	$-fc y'$	$fc x'$
f	ac	bc

Table 4: The partial derivatives of u and v with respect to each of the camera viewpoint parameters and the focal length according to our fully projective solution. Here $[a, b, c] = [x' + d'_x, y' + d'_y, \frac{1}{z' + d'_z}]$.

the translation vector is defined in the camera coordinate frame. Redefine:

$$[x, y, z]^T = \mathbf{R}\mathbf{p} + \mathbf{t}, \quad (12)$$

$$\text{then: } [d'_x, d'_y, d'_z]^T = \mathbf{t}, \quad (13)$$

and Eqs. (10) and (11) can be collapsed into:

$$[u, v] = f \left[\frac{x' + t_x}{z' + t_z}, \frac{y' + t_y}{z' + t_z} \right]. \quad (14)$$

In this case, the least-squares minimization procedure gives the estimates of the translation vector directly.

6 Experimental Results

In order to compare the three algorithms described in the previous sections we report extensive experiments with synthetic data. Our goal is to estimate the relative accuracy and convergence speed of each algorithm for a number of useful situations. So, in the tests we control a few parameters explicitly and sample all the others uniformly, hoping to cover important cases while keeping the amount of data down to a manageable level. In Lowe’s approximation, we use the depth of the center of the object in the camera frame as the multiplicative factor that yields the values of d_x and d_y . All the methods are tested with exactly the same poses and initial conditions [2].

Unless explicitly stated otherwise, all the experiments described here take the imaged object to be the eight corners of a cube, with edge lengths equal to

25 times the focal length of the camera (for a 20 mm lens, for instance, this corresponds to a half-meter-wide, long and deep object). The parameters explicitly controlled, in general, are the depth of the object’s center with respect to the camera frame (z_{true}), measured in focal lengths, and the magnitudes of the translation (t_{diff}) and the rotation (r_{diff}) needed to align the initial solution with the true pose. z_{true} is always measured in focal lengths and t_{diff} and r_{diff} are measured as a relative error with respect to z_{true} and as an absolute error in π radians, respectively. A formal definition of these parameters and of the whole sampling methodology is given in [2].

Unless stated otherwise, three average values are chosen for each of those parameters (Table 5). For each average value v , the corresponding parameter is then sampled uniformly in the region $[\frac{3v}{4}, \frac{5v}{4}]$.

Param	Avg 1	Avg 2	Avg 3
z_{true}	50	500	5,000
t_{diff}	0.1	0.01	0.001
r_{diff}	0.2	0.02	0.002

Table 5: General average sampling values used in most tests for the controlled parameters.

The other nine pose and initial solution parameters are in general sampled uniformly over their whole domain. The true object position is constrained to lie in the interior of the infinite pyramid whose origin is the optical center and whose faces are the semi-planes $z = |x|$ and $z = |y|$, $z \geq 0$.

For each test we compute two global image-space error measures, assuming known correspondence between image and model features. The first, called *Norm of Distances Error (NDE)*, is the norm of the vector of distances between the positions of the features in the actual image and the positions of the same features in the reprojected image generated by the estimated pose. The second, called *Maximum Distance Error (MDE)*, is the greatest absolute value of the vector of error distances. Both measures are always expressed using the focal length as length unit.

NDE and MDE do not necessarily indicate how close the estimated pose is from the true pose. We also record individual errors for six different pose parameters: the errors in the x , y and z coordinates

of the estimate for the actual object translation vector, measured as relative errors with respect to the object’s center actual depth (z_{true}), and the absolute errors in the estimates for the roll, pitch and yaw angles of the object frame with respect to the camera, measured in units of π radians. Although all these metrics were computed, this note usually shows only results with NDE, and x -translation error: they are faithfully representative of both image-space error metrics and the three translation and three rotation error metrics.

For each of these eight different error measures, we compute the average, the standard deviation, the averages and standard deviations excluding the 1%, 5% or 25% smallest and largest absolute values, and the median. Statistics that leave out the tails of the error distributions are included to be fair to a method (if any) that underperforms in a few exceptional situations but is better “in general”: for instance one that occasionally violently diverges but usually gives better results. In this note we usually present only the average error and its standard deviation and the results with the exclusion of the upper and lower 25% of the errors. For more error measures and more statistics see [2].

6.1 Convergence in the General Case

Initially, we tried to compare the speed of convergence and final accuracy of each method with arbitrary poses and initial conditions. The statistics for the NDE, based on 13,500 executions per method, are plotted in Fig. 1. They show that for most poses Lowe’s original approximation converges to a very high global error level and Ishii’s approximation only improves the initial solutions in its first iteration and diverges after that. Our fully projective solution, on the other hand, converges at a superexponential rate to an error level roughly equivalent to the relative rounding error of double precision, which is about 1.11×10^{-16} .

Even taking into account the worst data, our approximation still converges superexponentially to this maximum precision level — the bad cases only slow convergence a bit. But in this case Lowe’s original algorithm and (especially) Ishii’s approximation tend to diverge, yielding some solutions worse than the initial conditions.

The statistics for the errors in the individual pose parameters make the superiority of the fully projective approach even more clear. Fig. 2 exhibits the relative errors in the value of the x translation. Both Lowe’s and Ishii’s algorithms diverge in most situations, while the fully projective solution keeps its superexponential convergence. Due to their simplifications, Lowe’s and Ishii’s methods in those cases are not able to recover the true rotation of the object. They tend to make corrections in the translation components to fit the erroneously rotated models to the image in least-squares sense, generating very imprecise values for the parameters themselves. This problem is especially acute with Ishii’s approximation, which tends to translate the object as far away from the camera as possible, so that the re-projected images for all points are collapsed into a single spot that minimizes the mean of the squared distances with respect to the true images. Similar results were obtained for the other five parameter-space errors.

To assure that the results did not depend on symmetries in the cubical imaged object, we repeated the same tests with an asymmetric object whose eight points were all uniformly sampled in the space $[-1, 1]^3$ and then scaled for a maximum edge size of 25 focal lengths. All the results were almost identical to those obtained with the cube.

6.2 Convergence with Rough Alignment

For some relevant practical applications, our initial assumption that all the attitudes of the object with respect to the camera happen with equal probability is too general. For instance, in vehicle following applications it is reasonable to assume that the poses in which the object frame is roughly aligned to the camera frame happen with much larger probability than poses in which the object frame is rotated by large angles. We therefore performed some tests in which the rotation component of the initial solutions was represented by a quaternion whose axis was sampled uniformly on a unit semi-sphere with $z \geq 0$, but whose angle was constrained to the region $[-\frac{\pi}{5}, \frac{\pi}{5}]$.

The NDE statistics, plotted in Fig. 3, show that in this case the accuracy of Ishii’s approximation is much improved (predictably, given its semantics). Instead of diverging, now it converges exponentially to-

wards the rounding error lower bound. So, even in this favorable situation, Ishii’s approximation is still much less efficient than the fully projective solution, that converges super-exponentially (in about 5 iterations) for the NDE, as shown, and also for all other error metrics tested.

6.3 Execution Times

Lowe’s and Ishii’s simplifications do not result in a significant inner-loop performance gain with respect to the fully projective solution. We hand-optimized the three algorithms, with common subexpression factorization, loop vectorization and static pre-allocation of all matrices. After that, the internal loop (in Matlab) for Lowe’s method (which is the simplest of the three) contained only four floating point operations less than the internal loop of the fully projective solution.

We measured the execution times of 20 iterations of each method (details in [2]). The statistics shown in Fig. 4 were gathered from a set of 13,500 runs per method, performed with the same sampling techniques employed in the convergence experiments.

Fully projective solution average times are 2.99% to 4.21% longer than those of Lowe’s original method, but the standard deviations of the elapsed times for Lowe’s solution are between 6% and 130% bigger than those of the fully projective. Thus, the fully projective approach may be more suitable for hard real-time constraints, due to its smaller sensitivity to ill-conditioned configurations. The problem is that Lowe’s original method is much more likely to face singularity problems in the resolution of the system described in Eq. (2), resulting in the execution of slower built-in Matlab routines. The fully projective approach looks even better when compared to Ishii’s solution. The explanation is that a careful subexpression factorization can save us the work that Ishii’s simplifications are designed to save, so we pay no time penalty for a solution that is less sensitive to the proximity of singularities [2].

6.4 Sensitivity to Depth in Object Center Position

We also performed some experiments to check the sensitivity of the techniques to individual variations

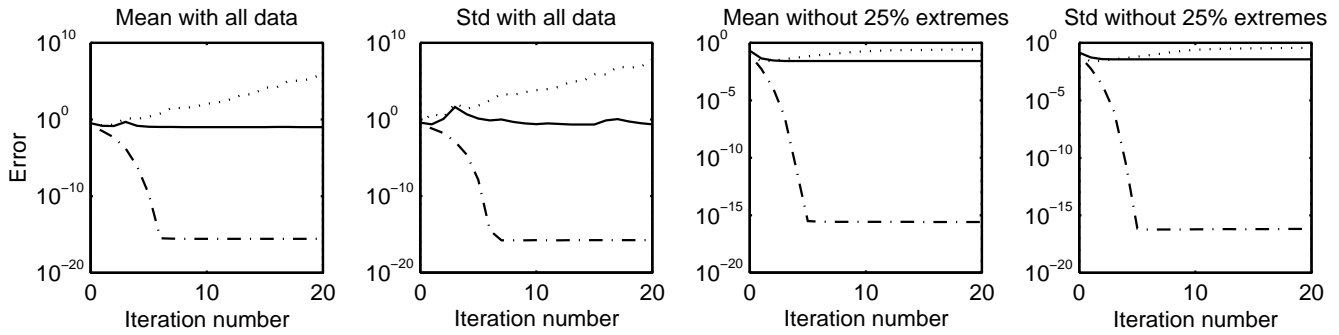


Figure 1: Convergence of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the number of iterations of Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line). Tests performed with a cube, rotated by arbitrary angles with respect to the camera frame.

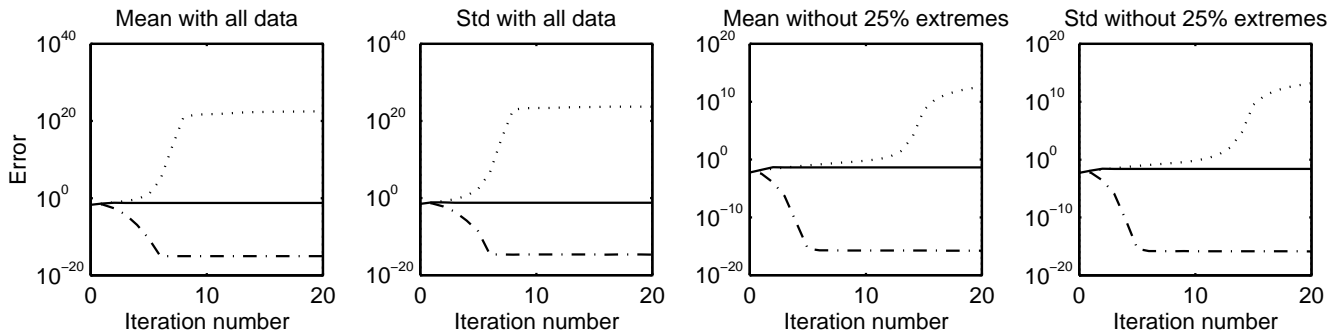


Figure 2: Convergence of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the number of iterations of Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line). Tests performed with a cube, rotated by arbitrary angles with respect to the camera frame.

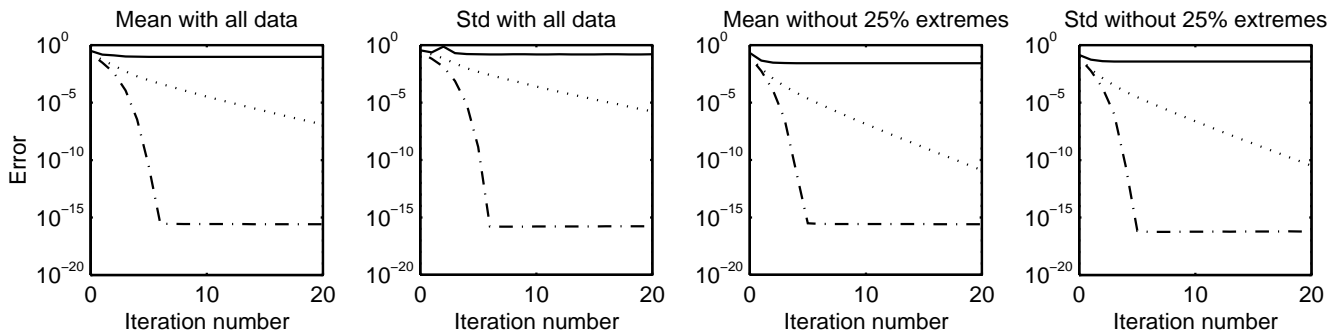


Figure 3: Convergence of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the number of iterations of Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line). Tests performed with a cube, rotated by angles of at most $\frac{\pi}{5}$ radians with respect to the camera frame.

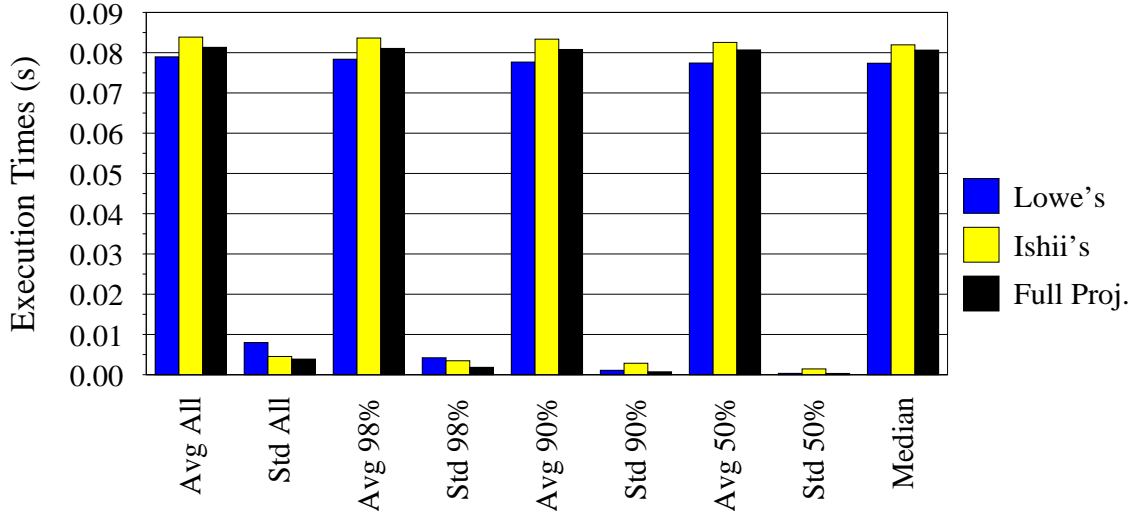


Figure 4: Execution times (in seconds) for 20 iterations of each method, computed over all data and with elimination of the 1%, 5% and 25% best and worst data.

in each one of the three controlled parameters. First, we varied the average value of z_{true} (object depth) logarithmically between 25 and 51,200 focal lengths, (corresponding, respectively, to 50 cm and 1,024 m, with a 20 mm lens). The statistics for the NDE, plotted in Fig. 5 for each of the twelve values chosen for z_{true} , show that our method is almost always much more accurate than both Lowe’s and Ishii’s. The only exception happens at the distance of 25 focal lengths.

The problem is that in this situation some individual object points may get as close as 5 focal lengths from the zero depth plane on the camera frame, due to the errors in the initial conditions. In this case, our method tends to behave like Ishii’s, shifting the object as far away from the camera as it can (so as to collapse the image in a single point), instead of aligning it. This can be confirmed by the analysis of the errors for the x translation (Fig. 6). But even in this extreme situation, our method, unlike Lowe’s and Ishii’s, still converges in most cases. The results for the errors on the rotation also support these observations.

6.5 Sensitivity to Translational Error in Initial Solution

Using the same sampling methodology as the previous experiment, we also studied the effect of changing the relative error in the translational component of the

initial pose estimates. Fifteen values for the relative initial translational error t_{diff} ranging from 0.025 to 0.5 were chosen.

The statistics for the NDE, depicted in Fig. 7, show that our method is once again much more accurate in general. However, when the average magnitude of the translational error is greater than 30% of the actual depth of the object’s center, our method has convergence problems for the worst 1% of the data and its overall reprojection accuracy drops to a level close to that of Lowe’s original approximation.

An analysis of the statistics for the x translation (Fig. 8 — other translation and pose angle results are similar) shows that in these cases no divergence towards infinite depth happens, but merely a premature convergence to false local minima. It is interesting to notice that the accuracy of Lowe’s method stays at this same high error levels even with much better initial conditions, which indicates that Lowe’s algorithm (as well as Ishii’s, which performs even worse) usually (and not only in extreme cases) gets stuck in local minima.

6.6 Sensitivity to Rotational Error in Initial Solution

Using the same sampling strategy once more, we selected ten average values for the absolute rotational error r_{diff} , ranging from $\frac{\pi}{10}$ to π radians. The statis-

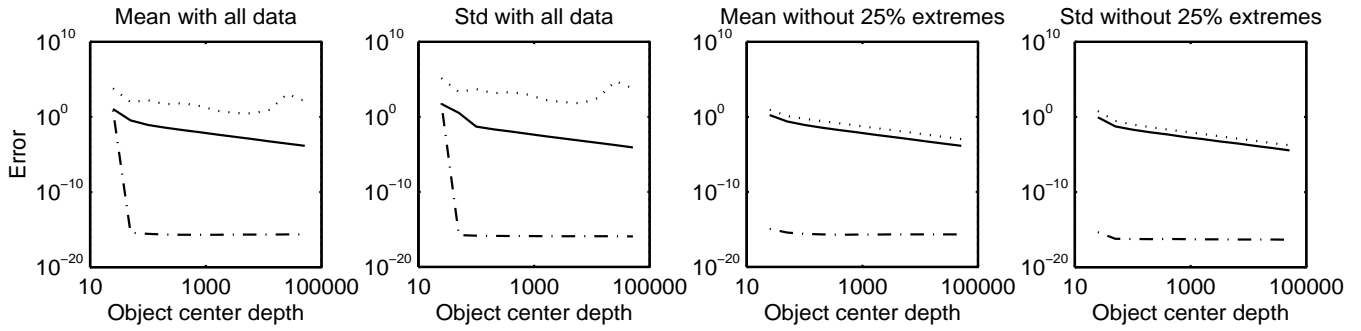


Figure 5: Sensitivity of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the actual depth of the object's center (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

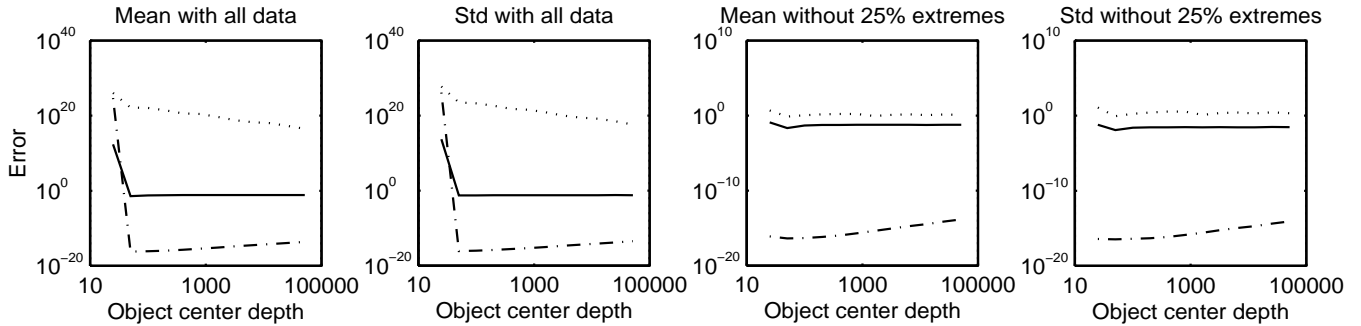


Figure 6: Sensitivity of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the actual depth of the object's center (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

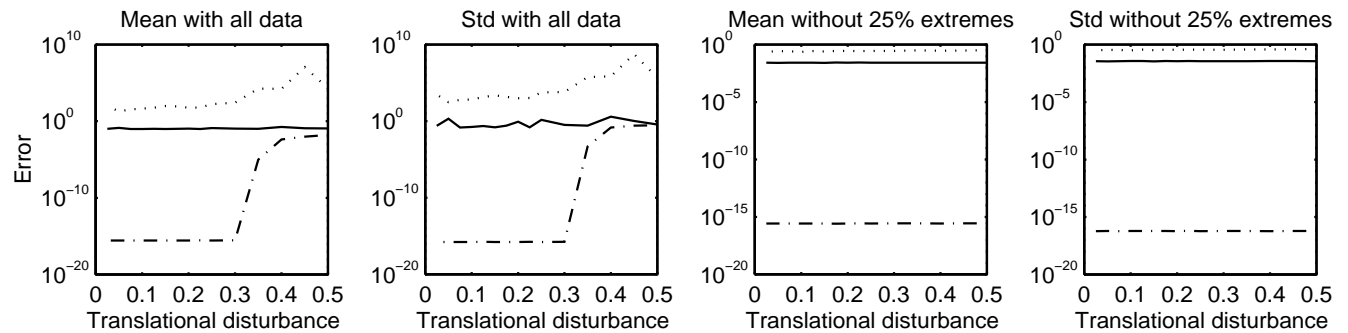


Figure 7: Sensitivity of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the ratio between the magnitude of the translational disturbance in the initial solution and the actual depth of the object's center, for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

tics for the NDE, exhibited in Fig. 9, show again the superiority of our approach for relatively small errors. Similarly, with errors bigger than $\frac{3\pi}{10}$ radians, our method starts having convergence problems and its reprojection accuracy approaches that of Lowe’s.

The errors in x translation recovery (Fig. 10) and in the pose angles show that large errors in initial rotation, unlike those in initial translation, make our method diverge towards infinite depth. This causes its accuracy in terms of pose parameters values to drop to levels comparable to (in some cases even worse than) those of Ishii’s. However, in this situation Lowe’s original method also diverges. A solution with a relative translational error of 10^{10} , 10^5 , or even 10^1 is not much more useful in practice than another solution with a relative translational error of 10^{20} . The problem in this case is the intrinsically downhill nature of Newton’s method, which is the core of all the techniques studied here. We believe that the only way to overcome this limitation would be to use a method based on an optimization technique with better global convergence properties, such as trust-region optimization.

6.7 Sensitivity to Additive Noise

In this experiment, Gaussian noise with zero mean and controlled standard deviation was added to the coordinates of the features in the image. 2,700 executions of each method were performed for each of the fifteen values of the noise standard deviation chosen in the range of 2^{-15} to 2^{-1} focal lengths.

The statistics for the NDE, plotted in Fig. 11, show that in this case the accuracy of our solution is always limited by the noise level, while the other two approaches get stuck on higher error levels even when the noise level is very small. For an error level of about 10^{-3} focal lengths (which corresponds roughly to the quantization noise with a sensing array of $1\text{K} \times 1\text{K}$ pixels), there is still a considerably wide gap of accuracy (about one order of magnitude) between our technique and Lowe’s, the second most accurate method.

The analysis of the effect on the x translation errors (Fig. 12) shows that divergence towards infinite depth is a problem again for relatively high noise levels (greater than 10^{-3} focal lengths in the worst cases). However, the roll angle errors, displayed in

Fig. 13, illustrate the fact that the degradation in the estimate for the rotation provided by our method happens smoothly. Our technique remains significantly more precise, at least for rotation recovery, for noise levels of up to 10^{-1} focal lengths. This is quite impressive given the fact that the restrictions in the view angle constrain the images to a 2×2 window (in focal lengths) on the image plane, where the noise was added.

6.8 Accuracy in Practice

Finally, we also wanted to compare the three methods in a realistic situation, in order to check if the better accuracy properties of our approach would make any difference in practice. The introduction of noise in the experiments was a first step towards this direction, but up to this point we have not addressed the question of what would be realistic initial conditions. One possibility for applications such as tracking would be to create reasonably precise initial estimates of the pose with a smoothing filter. But this approach is very dependent on application-specific parameters, such as the sampling rate of the camera, the bandwidth of the image processing system as a whole, the positional depth, the linear speed and the angular speed of the tracked object.

A more general approach, which we follow here, is to use a weaker camera model to generate an initial solution for the problem analytically, and then use the projective iterative solution(s) to refine this initial estimate. This approach was suggested by DeMenthon and Davis [4], who introduced a way of describing the discrepancy between a weak-perspective solution and the full-perspective pose with a set of parameters that can then be refined numerically, yielding the latter from the former. Let \mathbf{p}_i be the description of the i -th model point in the model frame and $[u_i, v_i]$ be the corresponding image, $1 \leq i < n$. Then, the weak-perspective solution proposed in that paper amounts to solving the following set of equations (in a least-squares sense), for the unknown three-dimensional vectors \mathbf{x} and \mathbf{y} :

$$\begin{aligned} (\mathbf{p}_i - \mathbf{p}_0) \cdot \mathbf{x} &= u_i - u_0, & 1 \leq i < n, \\ (\mathbf{p}_i - \mathbf{p}_0) \cdot \mathbf{y} &= v_i - v_0, & 1 \leq i < n. \end{aligned} \tag{15}$$

A normalization of these vectors yields the first two rows of the rotation component of the transformation

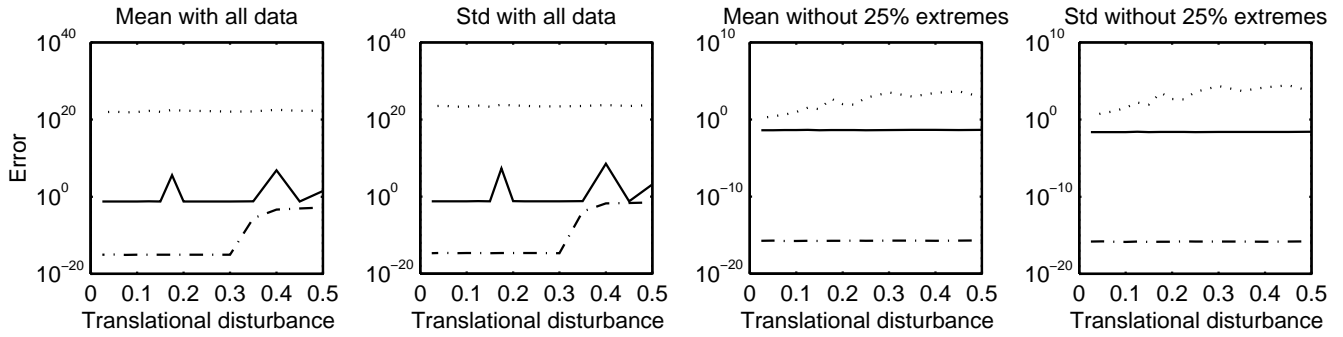


Figure 8: Sensitivity of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the ratio between the magnitude of the translational disturbance in the initial solution and the actual depth of the object's center, for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

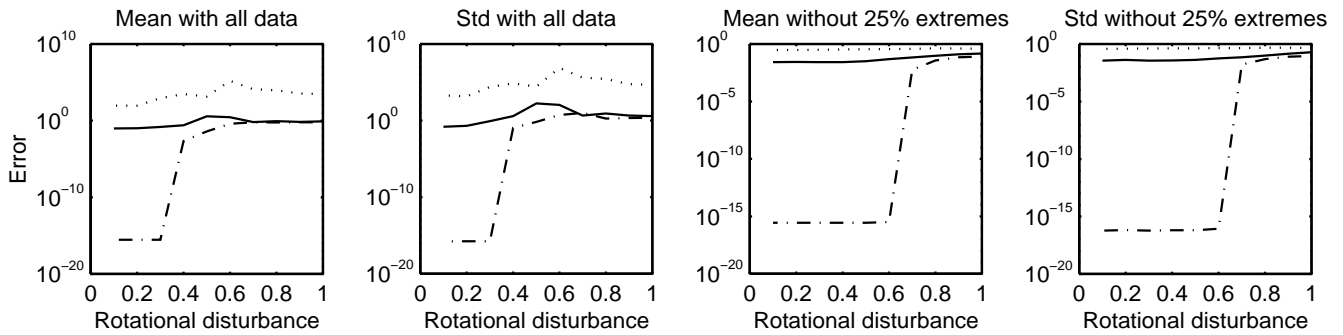


Figure 9: Sensitivity of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the magnitude of the rotational disturbance in the initial solution (in π radians), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

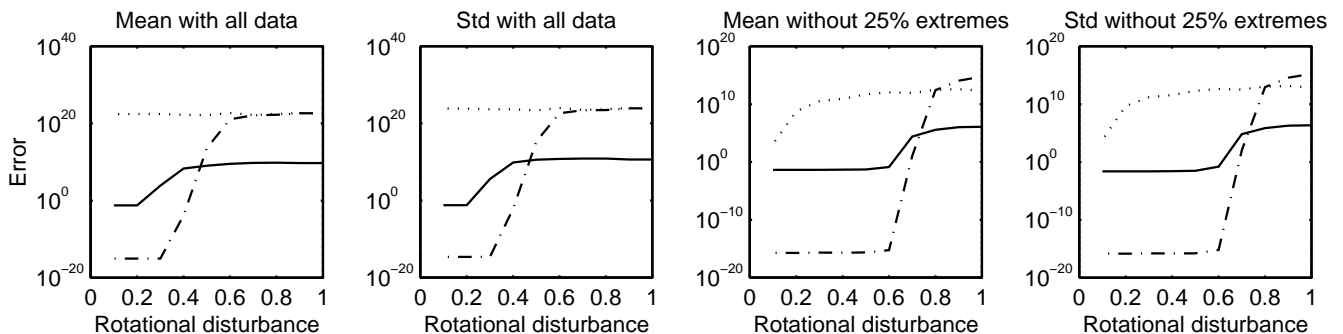


Figure 10: Sensitivity of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the magnitude of the rotational disturbance in the initial solution (in π radians), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

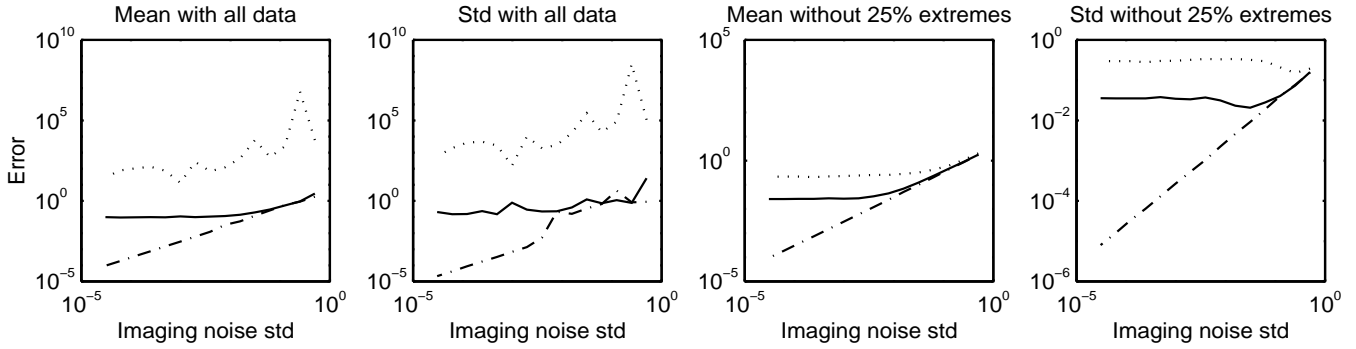


Figure 11: Sensitivity of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the standard deviation of the noise added to the image (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

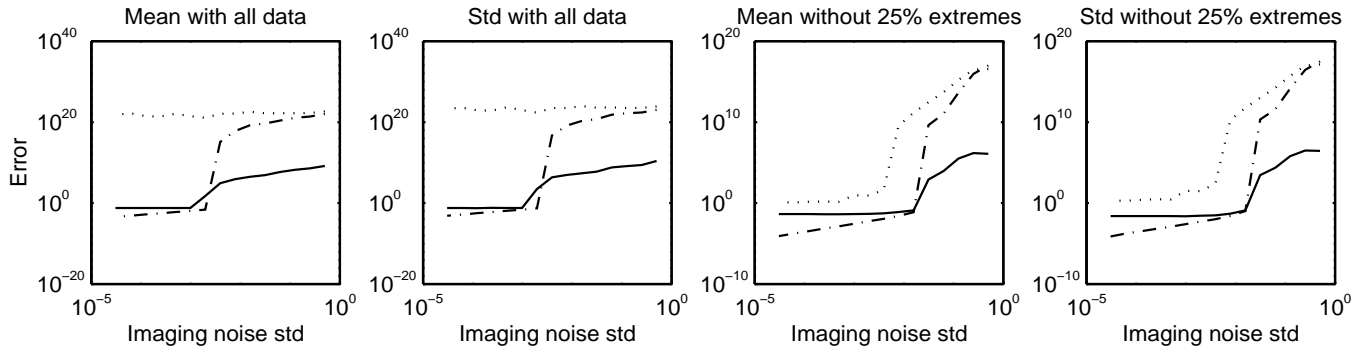


Figure 12: Sensitivity of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the standard deviation of the noise added to the image (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

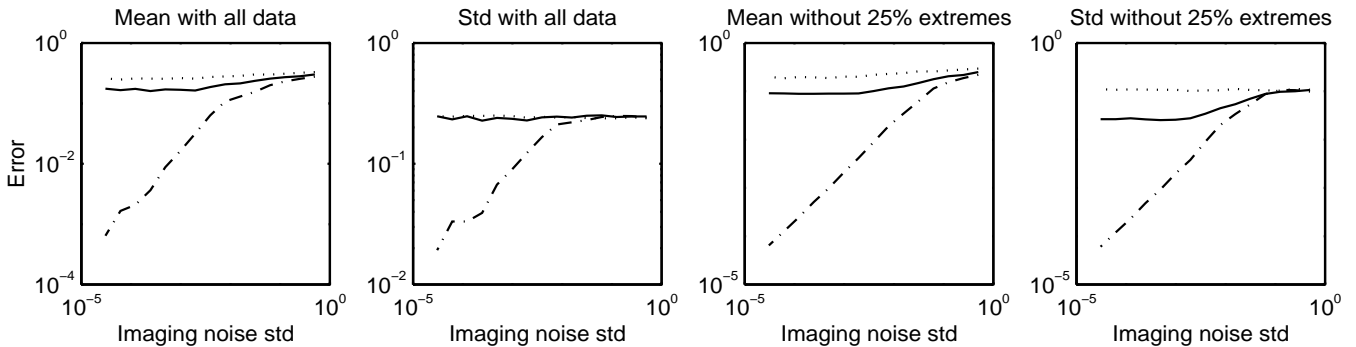


Figure 13: Sensitivity of the error on the estimated roll angle (measured in π radians), with respect to the standard deviation of the noise added to the image (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line).

that describes the object frame in the camera coordinate system. The third row can then be obtained with a single cross product operation. After that, the recovery of the translation is straightforward.

However, this simple weak-perspective approximation introduces errors that increase proportionally not only to the inverse depth of the object, but also to its “off-axis” angle (the angle of its center with respect to the optical axis, as viewed from the optical center). In order to avoid this last problem, we first preprocessed the image to simulate a rotation that puts the center of the object’s image in the intersection of the optical axis with the image plane. Let the center of the object image be described by $[u, v]$. Then, this transformation, as suggested in [22], is given by:

$$R = \begin{bmatrix} \frac{1}{d_1} & 0 & -\frac{u}{d_1} \\ \frac{uv}{d_1 d_2} & \frac{d_1}{d_2} & -\frac{v}{d_1 d_2} \\ \frac{u}{d_2} & \frac{v}{d_2} & \frac{1}{d_2} \end{bmatrix}, \text{ where:} \quad (16)$$

$$d_1 = \sqrt{u^2 + 1}, \quad d_2 = \sqrt{u^2 + v^2 + 1}.$$

After this preprocessing, we applied the technique described by Eq. (15), in order to recover the “foveated” pose. Then, we premultiplied the resulting transformation by the inverse of the matrix defined in Eq. (16), in order to recover the original weak-perspective pose, which was used as the initial solution for the iterative techniques being compared.

The only controlled parameter left was the actual depth of the object’s center (z_{true}). We chose nine average values for it, growing exponentially from 25 to 6,400 focal lengths. The noise standard deviation was set at 0.002 focal lengths (corresponding roughly to a 512×512 spatial quantization). The number of iterations of each method per run was set at 2, allowing a real-time execution rate of about 100 Hz. For each average value of z_{true} , 2,500 independent runs of each technique were performed.

The statistics for the NDE, depicted in Fig. 14, show that our fully projective solution was up to one order of magnitude more accurate than the other two methods for most cases in which the distance was smaller than 1,000 focal lengths (about 20 m, with the typical focal length of 20 mm). For distances bigger than that, the precision of the weak-perspective initial solution alone was bigger than the limitation

imposed by the noise and so the three techniques performed equally well.

Analysis of the results for the x translation error (Fig. 15) and the other five parameter-space errors, shows the interesting fact that all the techniques exhibit parameter-space accuracy peaks in the range of 50 to 400 focal lengths. The explanation for that is the fact that when the object gets too close, the quality of the initial weak-perspective solution degrades quickly. But on the other hand, when the object is too far away, the noise gradually overpowers the information about both the distance (via observed size) and the orientation of the object, since all the feature images tend to collapse into a single point. Of course, in practice, the exact location of these peaks depends on the dimensions of the actual object(s) whose pose is being recovered.

In the case of our technique, the accuracy peak happened clearly at distances of 50 to 100 focal lengths (1 to 2 m with 20 mm lens). Similar results were obtained when the number of iterations for each run was raised to 5. This suggests that our solution may be very well suited for indoor applications in which it is possible to keep a safe distance between the objects of interest and the camera.

7 Discussion and Conclusion

This note formulates a fully projective treatment of a pose- or parameter-recovery algorithm initially proposed by Lowe [13; 14; 15]. The resulting formulation is compared with formulations by Lowe and Ishii [11] that approximate the fully projective case. Many experiments based on different scenarios are presented here, and more are available in [2].

Lowe’s approximation was discussed by McIvor [16]. He states that assuming that d_x and d_y are constants amounts to an affine approximation. This is true for the parameters d_x and d_y themselves, but the affine approximation does not extend through the whole formulation — in Eq. (4) the denominators use $z' + d_z$ instead of just d_z . If a constant value had been used for those denominators then the formulation would be purely affine. Without implementing other formulations, McIvor speculates (correctly) that the use of full perspective would improve the accuracy of the viewpoint, perhaps at the expense

of decreased numerical stability. But as we show in Section 6, the fully projective formulation is actually **more** stable except in situations that break the other two formulations tested as well.

Bray [3] uses Lowe’s algorithm without discussing the approximation. Worrall *et al.* [21] compare their algorithm for perspective inversion with Lowe’s algorithm. They claim that their technique outperforms both Lowe’s original method and a reformulation of it using fully perspective projection, in terms of speed of convergence in simulations performed with a cube. This work sounds similar to ours, but [21] provides no detail on the perspective projection version of Lowe’s algorithm used in the comparison. They also do not present any discussion or comparison between the two different implementations of Lowe’s algorithm that they mention. Finally, they only report concrete experimental results for their own inversion method, which is based on line (rather than point) correspondences. No comparative evaluation of the two variants of Lowe’s algorithm was presented.

Our experiments indicate that a straightforward reformulation of the imaging equations removes mathematical approximations that limit the precision of Lowe’s and Ishii’s formulations. The fully projective algorithm has better accuracy with a minimal increase in terms of computational cost per iteration (Fig. 16).

The fully projective solution is very stable for a wide range of actual object poses and initial conditions. In some particularly extreme scenarios, our approach does suffer from numerical stability problems, but in these situations the accuracy of Lowe’s and Ishii’s approximations is also unacceptable, with errors of one or more orders of magnitude in the values of the pose parameters. We believe that this type of problem is a consequence of Newton’s method and can only be overcome with the use of more powerful numerical optimization techniques, such as trust-region methods.

In scenarios that may realistically arise in applications such as indoor navigation, with the use of reasonable (weak-perspective) initial solutions and taking into account the effect of additive Gaussian noise in the imaging process, the fully projective formulation outperforms both Lowe’s and Ishii’s approximations by up to an order of magnitude in terms of accuracy, with practically the same computational

cost.

References

- [1] M. A. Abidi and T. Chandra. A new efficient and direct solution for pose estimation using quadrangular targets: Algorithm and evaluation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(5):534–538, 1995.
- [2] H. Araujo, R. L. Carceroni, and C. M. Brown. A fully projective formulation for Lowe’s tracking algorithm. Technical Report 641, University of Rochester Computer Science Dept. Nov. 1996.
- [3] Alistair J. Bray. Tracking objects using image disparities. *Image and Vision Computing*, 8(1):4–9, 1990.
- [4] D. F. DeMenthon and L. S. Davis. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15:123–141, 1995.
- [5] M. Dhome, M. Richetin, J-T. Lapresté, and G. Rives. Determination of the attitude of 3-D objects from a single perspective view. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(12):1265–1278, 1989.
- [6] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [7] Donald B. Gennery. Visual tracking of known three-dimensional objects. *International Journal of Computer Vision*, 7(3):243–270, 1992.
- [8] R. M. Haralick and C. Lee. Analysis and solutions of the three point perspective pose estimation problem. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 592–598, 1991.
- [9] R. Horaud, S. Christy, F. Dornaika, and B. Lamiroy. Object pose: Links between paraperspective and perspective. In *Proc. 5th IEEE International Conference on Computer Vision*, pages 426–433, 1995.

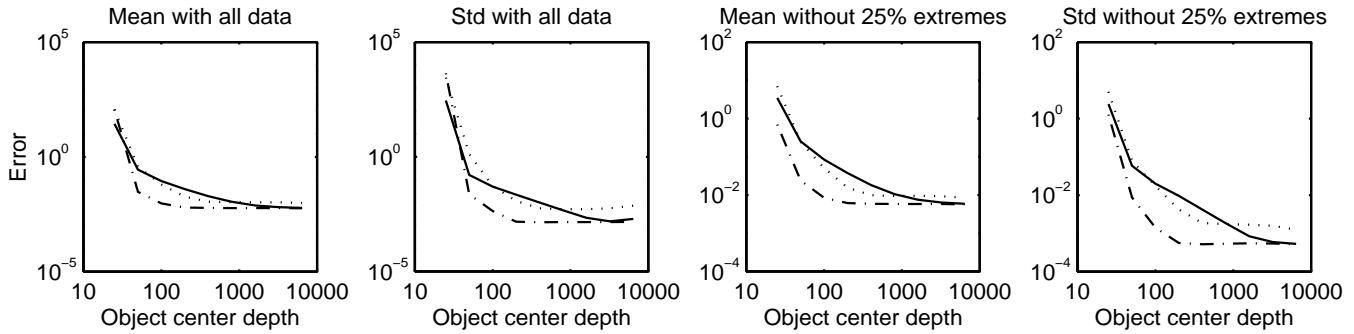


Figure 14: Sensitivity of an image-space error metric, the Norm of Distances Error (see introduction of Section 6), with respect to the actual depth of the object's center (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line). Tests performed with initial solutions generated by a weak-perspective approximation.

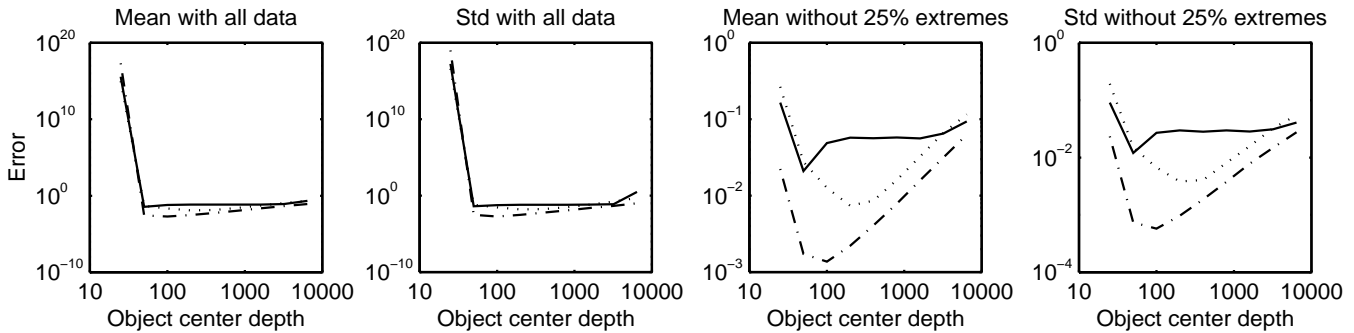


Figure 15: Sensitivity of the ratio between the error on the estimated x translation and the actual depth of the object's center, with respect to the actual depth of the object's center (in focal lengths), for Lowe's (solid line), Ishii's (dotted line), and our fully projective solution (dash-dotted line). Tests performed with initial solutions generated by a weak-perspective approximation.

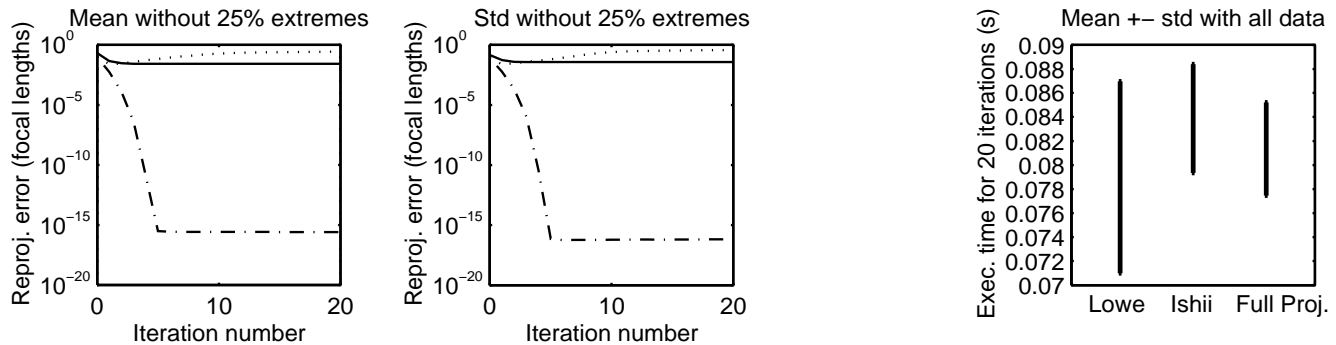


Figure 16: Summary. Left (subset of Fig. 1): convergence of the NDE, an image-space error metric (see Section 6), with respect to the number of iterations of Lowe's (solid line), Ishii's (dotted line), and the fully projective solution (dash-dotted line); statistics exclude the best and worst 25% results. Right (subset of Fig. 4): mean and standard deviation of execution times; statistics include all data.

- [10] R. Horaud, B. Conio, O. Le Boulleux, and B. Lacroix. An analytic solution for the perspective 4-point problem. *Computer Vision, Graphics, and Image Processing*, 47:33–44, 1989.
- [11] M. Ishii, S. Sakane, M. Kakikura, and Y. Mikami. A 3-D sensor system for teaching robot paths and environments. *International Journal of Robotics Research*, 6(2):45–59, 1987.
- [12] S. Linnainmaa, D. Harwood, and L. S. Davis. Pose determination of a three-dimensional object using triangle pairs. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(5):634–647, 1988.
- [13] David G. Lowe. Solving for the parameters of object models from image descriptions. In *Proc. ARPA Image Understanding Workshop*, pages 121–127, 1980.
- [14] David G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.
- [15] David G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.
- [16] Alan McIvor. An analysis of Lowe’s model-based vision system. In *Proc. 4th Alvey Vision Conference*, pages 73–77, University of Manchester, U.K., 1988.
- [17] N. Navab and O. Faugeras. Monocular pose determination from lines: Critical sets and maximum number of solutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–260, 1993.
- [18] D. Oberkampf, D. F. DeMenthon, and L. S. Davis. Iterative pose estimation using coplanar feature points. *Computer Vision and Image Understanding*, 63(3):495–511, 1996.
- [19] T. Q. Phong, R. Horaud, and P. D. Tao. Object pose from 2-D to 3-D point and line correspondences. *International Journal of Computer Vision*, 15:225–243, 1995.
- [20] T. Shakunaga and H. Kaneko. Perspective angle transform: Principle of shape from angles. *International Journal of Computer Vision*, 3:239–254, 1989.
- [21] A. D. Worrall, K. D. Baker, and G. D. Sullivan. Model based perspective inversion. *Image and Vision Computing*, 7(1):17–23, 1989.
- [22] Y. Wu, S. S. Iyengar, R. Jain, and S. Bose. A new generalized computational framework for finding object orientation using perspective trihedral angle constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(10):961–975, 1994.