

Pose Estimation using L_∞

X. Zhang¹

¹ University of Sydney, Dept. Computer Science

Email: xianhang@cs.usyd.edu.au

Abstract

A method is presented for estimating the pose (position & orientation) of a camera based on a set of correspondences between world space and camera space by decomposing the problem into orientation estimation and triangulation. Both of these steps can be proven to be globally convergent which allows fast and accurate estimation. Furthermore, the pose estimation works in full projective space, preserves the full orthonormality of the rotation matrix does not make any assumptions about the co-planarity of the points and provides an elegant method of doing outlier removal. Experiments show that the presented method of pose estimation is more accurate than currently available schemes and is fast enough to run in real time.

Keywords: Pose Estimation, Perspective-n-point, Triangulation, Augmented Reality, L_∞

1 Introduction

The computation of the pose (position and orientation) of a camera is known as pose estimation. When this estimation is computed based on correspondences between points in an image and points in the scene whose locations are known, the problem is termed the *Perspective-n-point* problem. Such a problem is important in fields such as Augmented Reality in which the position of the head mounted display must be known to overlay virtual data over the scene as well as other fields such as robotics and photogrammetry.

There are 4 desirable properties which we would like of a pose estimation algorithm (See **Table 1** for summary). Firstly, it is desirable for the quality of the pose estimation to be globally optimal and not depend on a prior estimate. For some iterative style pose estimation algorithms, the choice of an initial estimate is very important as the algorithm can be trapped in a local minima. Secondly, we would like the algorithm to handle coplanar and non-coplanar points equally well. Many algebraic based solutions have trouble handling coplanar points since they produce degenerate solutions. Thirdly, we would like the algorithm that returns an orthonormal rotation matrix. Even though the rotation matrix has 9 variables, actual rotation only has 3 degrees of freedom and is, thus, over constrained. Ideally, we would like an estimate which preserves the orthonormal nature of the rotation matrix. Finally, the choice of camera model used by the estimation algorithm is important. Many iterative style algorithms work using a simplified version of the perspective camera model

such as Affine, Projective, Scaled Orthographic Perspective or Weak Perspective.

Closed form solutions can be computed in projective space for between 3 and 6, non-coplanar points [1], [2], [3], [4], [5], [6], [7], [8], [9], however, for more than 6 points, no exact closed form solutions exist. Only the 3 point solution preserves full orthonormality. Some of the earliest iterative pose estimates relied on a Newton-Raphson style method and used an affine camera model which is a highly simplified model of a true perspective camera[10]. Later work has extended the concept to more accurate approximations as well as more sophisticated converging but all of these methods require an initial estimate and are not guaranteed to converge to the optimal value[11, 12, 13, 14]. [5] presents an alternative formulation of the problem called POSIT using a Scaled Orthographic Projection (SOP) camera model which does not require an initial estimate but does not fully preserve orthonormality. Furthermore, POSIT does not always converge onto the correct value. In certain conditions, POSIT can converge onto a completely wrong solution. [16, 17] present a method globally convergent algorithm that preserves full orthonormality and works using a Weak perspective model.

In this paper, an iterative pose estimation algorithm termed LIPE is presented which solves the pose estimation problem. LIPE works by decomposing the problem into first estimating the position of the camera for a known orientation using triangulation and then estimating orientation as a separate step. LIPE preserves the full orthonormal nature of the rotation matrix, handles co-planar and non-coplanar points equally well and utilises a full perspective

Table 1: A comparison of various Pose Estimation Algorithms

Method	Globally Optimal	Orthonormal Rotation Matrix	Camera Model	Handles Coplanar Points
Closed Form Solutions	Yes	No (except for 3 point)	Perspective	No
Lowe	No	Yes	Affine	Yes
Araujo	No	Yes	Projective	Yes
POSIT	Mostly	No	Scaled Orthographic Projection	No
Lu	Yes	Yes	Weak Perspective	Yes
LIPE	Probably	Yes	Perspective	Yes

model. LIPE also appears to be globally optimal based on experimental evidence but this is not mathematically proven. Furthermore, LIPE also presents an integrated and elegant way to do outlier removal and performs more accurately than current algorithms on synthetic data.

The remainder of the article is organised as follows: In Section 2, a description of the LIPE algorithm is presented. In Section 2.1, a proof is provided that the LIPE algorithm is globally convergent. In Section 2.2, a method of doing outlier removal using LIPE is presented. In Section 3, a comparison between LIPE and other pose estimation algorithms is presented. Finally, Section 4 presents some conclusions.

2 LIPE

L-Infinity (L_∞) based Pose Estimation (LIPE) decomposes the pose estimation into two steps, orientation estimation and triangulation. If we assume that we know the orientation matrix R for a calibrated, perspective camera, then we can formulate a correspondence between an image point x and world point X as:

$$C + \lambda x' = X, \lambda > 0, x' = Rx \quad (1)$$

Where x' is the vector x in world space co-ordinates. That is, a ray passing from the camera centre C in the direction x' will intersect the point X .

However, it is also equally accurate to write:

$$C = X - \lambda x', \lambda > 0, x' = Rx \quad (2)$$

That is, a ray passing from the point X in the opposite direction to x' will pass through the camera centre. Thus, if we have multiple image points x_i observing multiple world points X_i as in the Perspective-n-point problem, we would have the equation:

$$C = X_i - \lambda_i x'_i, \lambda > 0, i = 0, 1, \dots, n \quad (3)$$

In other words, the camera centre lies at the intersection of the rays back projected from the points.

In the absence of noise, all of these rays would meet at a single point and this equation would be trivial to solve. However, if noise is introduced, then no exact solution exists and an estimate is necessary. This problem then looks remarkably similar in structure to the Triangulation problem[18].

In Triangulation, we have multiple cameras of a known pose each of which observes a single point of unknown position. Each camera projects out a ray that the point lies on and multiple observations should produce multiple rays intersecting at a single point in the absence of noise or near a single point with noise. Instead, in this instance, we have multiple points of known position “observing” the position of a single camera.

[19] presents a solution to the Triangulation problem called L_∞ that works on the basis of a novel cost function. Rather than trying to use least squares (L_2) as a cost function, the L_∞ cost function is used instead. If we define the estimation error between x_i and X_i as ε_i , then the L_∞ cost function is:

$$L_\infty(C) = \max(|\varepsilon_0|, |\varepsilon_1|, \dots, |\varepsilon_n|) \quad (4)$$

As long as we restrict the search space to a convex domain in front of all cameras/points (ie: $\lambda_i > 0$), then the L_∞ cost function can be proven to have a single, global minima over C . Using the L_∞ algorithm, we can construct an estimate of the camera centre C for any given R . Thus, the problem becomes finding an estimate of the optimal R .

There are many alternative methods for estimating R and several existing pose estimation algorithms are capable of determining just rotation as a first step. However, in LIPE, the same L_∞ cost function can be also used for estimating R . For a given R , we can define:

$$L_\infty(R) = \min(L_\infty(C)) \text{ for all } C, \lambda > 0 \quad (5)$$

Then, estimating R is simply a matter of minimising $L_\infty(R)$.

2.1 Global Convergence in R

The LIPE cost function appears to be globally convergent in R although a proof is not provided. In experimental trials on synthetic data, LIPE has always found the global minima reliably in over 1,000,000 trials. LIPE has been implemented in a real-world augmented reality system with no indication that it has trouble in finding the global minima. It would appear that, under normal operating conditions, LIPE performs sufficiently reliably such that it can be regarded as globally convergent even if it may fail under degenerate conditions.

2.2 Outlier Removal

When doing L_∞ minimization, a frontier *set* of rays is kept track. A ray n is in the frontier set if:

$$|\varepsilon_n| = L_\infty(R)$$

Thus, those rays on the frontier represent the observations that disagree the most with the estimate of LIPE. Since outliers by definition lie far away from the bulk of the measurements, if outliers exist, then they must lie on the frontier. Outliers can be removed by successively removing rays from the frontier set and testing to see if the LIPE cost function has decreased. The removal of an outlier should significantly decrease the cost function while the removal of an inliers should only decrease it by a marginal amount. There are a number of advantage to using L_∞ to remove outliers rather than standard approaches such as RANSAC[20]. Firstly it is much faster to run. Since the frontier needs to be calculated for L_∞ anyway, there is almost no performance penalty in using the frontier for outlier removal.

Secondly it does not suffer from the mistake of fitting to outliers like L_2 . With L_2 outlier removal, if there are a large number of outliers or the outliers have a large amount of error, it is possible for L_2 to fit a estimate based on the outliers rather than the inliers. If this occurs, then inliers will be misclassified as outliers and removed until only outliers remain and a false fit is made. With L_∞ , since we are taking the maximum error rather than the weighted error, large numbers or excessively erroneous outliers to do not disproportionately affect the result.

3 Results

3.1 Accuracy

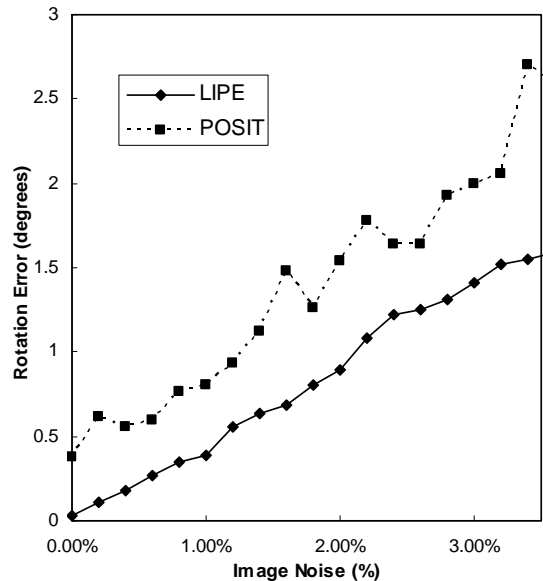


Figure 1: The average rotation error for both LIPE and POSIT in the presence of image noise.

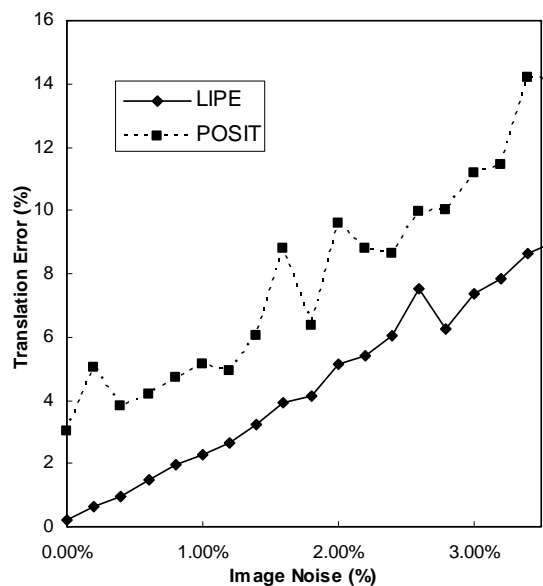


Figure 2: The average translation error for both LIPE and POSIT in the presence of image noise.

A synthetic benchmark of the LIPE algorithm was performed against POSIT which represents the current most popular iterative pose estimation algorithm. In software, a simulated random point cloud of 1000 landmarks were distributed within a 1unit by 1unit cube. A virtual camera was placed in the scene with a random rotation and displacement and 10 random correspondences were picked. Varying levels of image noise were introduced into the correspondence in intervals of 0.2% and each interval consisted of 100 trials which were averaged.

As can be seen from Figures 1 & 2, LIPE performs significantly more accurately than POSIT for a given amount of noise. This effect is especially pronounced in cases where the image noise is very low. The probable cause of this is due to the different camera models used. Scaled Orthographic Projection does not fully approximate the Perspective camera model, there is some base level of error in POSIT even in situations with no noise.

Although it is not apparent from the figures, in synthetic benchmarks, POSIT falsely converges onto completely wrong values somewhere between 3-10% of the time. These false convergences have not been taken into account when calculating the error values but would present a problem in real-world systems for which ground truth is not known.

3.2 Performance

Even though LIPE does not strictly need an initial estimate to obtain the correct answer, a good estimate allows it to converge much faster. In synthetic tests, LIPE performed roughly about an order of magnitude slower than POSIT given no initial estimate and about comparable to POSIT given an estimate with an error of around 10° . A real world system was implemented using LIPE in which the pose of a handheld webcam was determined in real time in a natural environment. The number of points tracked varied between 5 and 20 and outliers occurred roughly about 10% of the time. Because there are physical constraints on how fast a camera can move, LIPE was able to run very fast since the estimated camera pose does not change much from frame to frame. Running on a 1.6Ghz machine, each LIPE iteration took roughly 1- 3ms which indicates that it is suitable for real time operation.

4 Conclusions

A novel pose estimation algorithm called LIPE is presented that contains a number of attractive properties. Unlike previous iterative solutions, LIPE works using a full perspective camera model and does not require an initial estimate. Also, unlike many closed form solutions, LIPE handles co-planar and non-coplanar points equally well. Such a feature is important in many practical systems working in indoor environments which contain many large, planar surfaces.

LIPE also performs significantly more accurately than current popular algorithms and, given a good initial estimate, takes no longer to run. Hence, LIPE is well suited to indoor environments in which many points found will be co-planar and also in environments in which noise is low. This makes LIPE ideally suited to Augmented Reality scenarios in which landmarks can be determined to a high degree of accuracy beforehand and the pose estimate needs to be accurate

to prevent “jitter” in the image which can ruin the illusion. LIPE would also be useful in situations in which outliers have not been pruned beforehand.

A further avenue to explore would be to either mathematically prove or disprove the global convergence property of LIPE.

5 References

- [1] Y.I. Abdel-Aziz and H.M. Karara, “Direct Linear Transformation into Object Space Coordinates in Close-Range Photogrammetry,” *Proceedings of the Symposium of Close-Range Photogrammetry*, pp. 1-18, Jan. 1971.
- [2] D. DeMenthon and L.S. Davis, “Exact and Approximate Solutions of the Perspective-Three-Point Problem,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 11, pp. 1,100-1,105, Nov. 1992.
- [3] M. Dhome, M. Richetin, J. Lapreste, and G. Rives, “Determination of the Attitude of 3D Objects from a Single Perspective View,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, pp. 1,265-1,278, Dec. 1989.
- [4] R.M. Haralick, C. Lee, K. Ottenberg, and M. Nolle, “Analysis and Solutions of the Three Point Perspective Pose Estimation Problem,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 592-598, 1991.
- [5] R.M. Haralick and L.G. Shapiro, *Computer and Robot Vision*. Reading, Mass.: Addison-Wesley, 1993.
- [6] R. Horaud, B. Canio, and O. Le Boulleux, “An Analytic Solution for the Perspective 4-Point Problem,” *Computer Vision, Graphics, and Image Processing*, no. 1, pp. 33-44, 1989.
- [7] B.K.P. Horn, H.M. Hilden, and S. Negahdaripour, “Closed-Form Solution of Absolute Orientation Using Orthonormal Matrices,” *Journal of the Optical Society of America.*, vol. 5, pp. 1,127-1,135, 1988.
- [8] G.H. Rosenfield, “The Problem of Exterior Orientation in Photogrammetry,” *Photogrammetric Engineering.*, pp. 536-553, 1959.
- [9] E.H. Tompson, “The Projective Theory of Relative Orientation,” *Photogrammetria*, pp. 67-75, 1968.
- [10] D.G. Lowe, “Three-Dimensional Object Recognition from Single Two-Dimensional Image,” *Artificial Intelligence*, vol. 31, pp. 355- 395, 1987.

- [11] T.D. Alter, "3D Pose from Corresponding Points under Weak- Perspective Projection," Technical Report A.I. Memo No. 1,378, MIT Artificial Intelligence Lab., 1992.
- [12] H. Araujo, R. Carceroni, and C. Brown, "A Fully Projective Formulation for Lowe's Tracking Algorithm," Technical Report 641, Univ. of Rochester, 1996
- [13] R. Horaud, S. Christy, and F. Dornaika, "Object Pose: The Link between Weak Perspective, Para Perspective and Full Perspective," Technical Report RR-2356, INRIA, Sept. 1994.
- [14] D.P. Huttenlocher and S. Ullman, "Recognizing Solid Objects by Alignment with an Image," *International Journal on Computer Vision*, vol. 5, no. 2, pp. 195-212, 1990.
- [15] D. DeMenthon and L. Davis, "Model-Based Object Pose in 25 Lines of Code," *International Journal on Computer Vision*, vol. 15, pp. 123-141, June 1995.
- [16] R.M. Haralick et al., "Pose Estimation from Corresponding Point Data," *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1,426-1,446, 1989.
- [17] C. Lu, G. Hagar, E.Mjolsness, "Fast and Globally Convergent Pose Estimation from Video Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 22, no 6, pp. 610-622, June 2000.
- [18] R. I. Hartley and P. Sturm. "Triangulation", *Computer Vision and Image Understanding*, no 68(2), pp. 146–157, November 1997.
- [19] R. I. Hartley and F. Schaffalitzky, " L_∞ minimization in geometric reconstruction problems". *Conference of Computer Vision and Pattern Recognition*, volume I, pp. 504–509, Washington DC, USA, 2004.
- [20] M. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting and Automatic Cartography" *Communications of the ACM*, no. 6, pp. 381-395, 1981.