## Visual SFM-Based Measure

This measure computes camera motion given any pair of images from which at least eight matching point features can be extracted. It solves for the extrinsic transformations between camera poses (egomotion) up to a translation scale factor, and it also computes Structure From Motion (SFM) for each observed feature pair. The transformations between camera poses are then stored in order to evaluate subsequent trajectory hypotheses relative to them.

***SFM Algorithm:*** SFM is a very active field of research with a large variety of SFM algorithms existing in the literature. We employ a standard and state-of-the-art method similar to [3], [6] and [7]. The algorithm begins by finding SURF interest points in a pair of images. Matching points are identified by extracting a SURF feature vector at each point and evaluating pairs using the nearest neighbor criterion [4].

Following the development in [3], RANSAC [5] detects and removes matches that are not mutually consistent with a set of known intrinsic calibration parameters while simultaneously fitting a geometric model to the data. This process iteratively employs the eight-point algorithm [1] to estimate the essential matrix in a linear least squares sense.

The linear estimate of the essential matrix is then further refined as follows: The essential matrix is factored into the product of a skew symmetric matrix and a rotation matrix, resulting in four possible solutions for the extrinsic transformation. Then, each pair of matched points is triangulated using each of the four solutions, and the solution which results in a positive depth for the highest number of pairs is selected. This solution is then recomposed into a projection matrix that becomes the initial condition for Sparse Bundle Adjustment (SBA) [2], which refines the estimate via local nonlinear optimization.

***Eight-Point Algorithm:*** Given a pair of matched interest points  in homogenous image coordinates observed at two times  and , the geometric epipolar constraint can be defined as

,

where  is the calibration matrix containing the camera intrinsic parameters, and  is the essential matrix.

To obtain a linear estimate of the elements of the essential matrix, this equation can be rearranged as

,

where is a column vector made up of the entries of in row-major order, and  is a row vector made up of the remaining values. Each pair of matched points gives rise to one equation. Therefore, the linear least squares estimate of  is the null space of the matrix formed by stacking at least eight constraint equations in rows.

The numerical stability of this method can be improved by forcing the essential matrix to have exactly two equal singular values. This can be done by replacing its singular values as follows

,

where and are the orthonormal matrices obtained by singular value decomposition (i.e).

***Proposed Measure:*** Assuming that the camera frame and body frame are coincident, the essential matrix is related to the body motion as follows

,

where  converts rotations from quaternion to matrix form, and the bracket notation  converts its argument to the matrix operator form of the cross product.

DAVID WILL REWRITE THIS PARAGRAPH: Therefore, the recovered motion can be directly compared to the candidate trajectory which is represented as a relative quaternion and a relative translation. Hence, we compute the residual of as

.

References:

[1] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 19(6):580 - 593, October 1997.

[2] M.I.A. Lourakis and A.A. Argyros, “SBA: A Software Package for Generic Sparse Bundle Adjustment”, ACM Transactions on Mathematical Software, vol. 36, no. 1, pp. 1-30 (Mar. 2009).

[3] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", 2004.

[4] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), 2008.

[5] Martin A. Fischler and Robert C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". Comm. of the ACM, 1981.

[6] Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R.. “Visual modeling with a hand-held camera”. Int. J. of Computer Vision. 2004.

[7] Noah Snavely, Steven M. Seitz, Richard Szeliski, "[Photo tourism: Exploring photo collections in 3D](http://phototour.cs.washington.edu/Photo_Tourism.pdf)," ACM Transactions on Graphics 2006.

## Modeling the Residual

In order to obtain a better data-driven residual representation, the residual  resulting from the discussed measures is compared to its characteristic distribution as determined from training data. For instance, we have modeled the tracking error in the Middleburry dataset using a single Gaussian, a Gaussian Mixture Model, and Kernel Density Estimation (KDE). In order to select the most appropriate model, we divided our dataset into two sets for training and testing. The training data is used in the model learning, and then the model response against the testing data is compared to the actual groundtruth response using the sum of squared error (SSE). Figure 1 illustrates the obtained results. In our experiments, we found that even as simple as a single Gaussian is appropriate for modeling the error, and that the SSE decreases with increasing the number of Gaussian components, while being generally close to the SSE obtained when using the KDE. In our framework, we use a single Gaussian component in order to avoid over-fitting. Given the residual  for each feature pair and the error model (eg.), we compute the normalized Sum of Squared Differences (SSD)

.

As discussed in the previous section, follows a chi-square distribution; therefore, the final data likelihood is given as

.

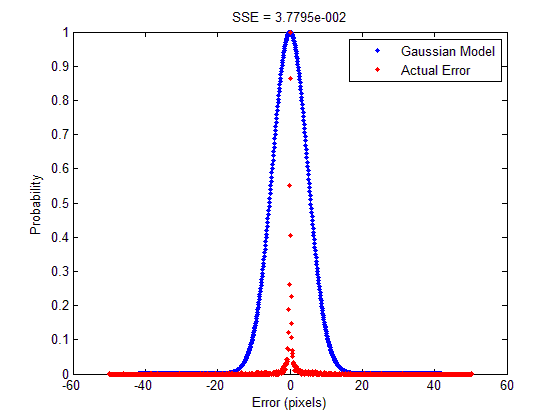


Figure . KLT feature tracking error distribution given Middlebury Art data. The blue curve represents the learned model, and the red dots represent the actual data.