Fast, Robust, Continuous Monocular Egomotion Computation

Andrew Jaegle*, Stephen Phillips*, and Kostas Daniilidis
University of Pennsylvania
Philadelphia, PA USA
{ajaegle, stephi, kostas}@seas.upenn.edu

***Abstract*— We propose robust methods for estimating camera egomotion in noisy, real-world monocular image sequences in the general case of unknown observer rotation and translation with two views and a small baseline. This is a difficult problem because of the nonconvex cost function of the perspective camera motion equation and because of non-Gaussian noise arising from noisy optical flow estimates and scene non-rigidity. To address this problem, we introduce the expected residual likelihood method (ERL), which estimates confidence weights for noisy optical flow data using likelihood distributions of the residuals of the flow field under a range of counterfactual model parameters. We show that ERL is effective at identifying outliers and recovering appropriate confidence weights in many settings. We compare ERL to a novel formulation of the perspective camera motion equation using a lifted kernel, a recently proposed optimization framework for joint parameter and confidence weight estimation with good empirical properties. We incorporate these strategies into a motion estimation pipeline that avoids falling into local minima. We find that ERL outperforms the lifted kernel method and baseline monocular egomotion estimation strategies on the challenging KITTI dataset, while adding almost no runtime cost over baseline egomotion methods.**

I. Introduction

Visual odometry in real-world situations has attracted increased attention in the past few years in large part because of its applications in robotics domains such as autonomous driving and unmanned aerial vehicle (UAV) navigation. Stereo odometry and simultaneous localization and mapping (SLAM) methods using recently introduced depth sensors have made dramatic progress on real-world datasets. Significant advances have also been achieved in the case of *monocular* visual odometry when combined with inertial information.

State-of-the-art visual odometry uses either the discrete epipolar constraint to validate feature correspondences and compute inter-frame motion [1] or directly estimates 3D motion and 3D map alignment from image intensities [2]. In contrast to the state of the art, in this paper we revisit the *continuous* formulation of structure from motion (SfM), which computes the translational and rotational velocities and depths up to a scale from optical flow measurements. Our motivation lies in several observations:

- UAV control schemes often need to estimate the translational velocity, which is frequently done using a combination of monocular egomotion computations and single-point depths from sonar [3].

*Authors contributed equally.

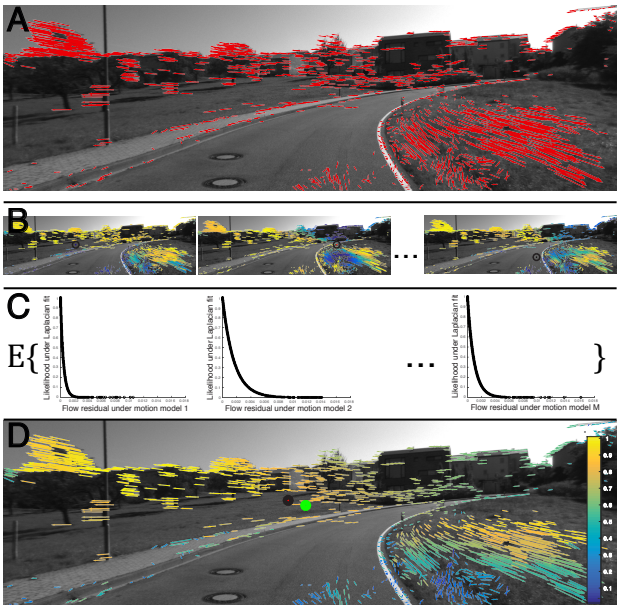

Fig. 1. Schematic depiction of the ERL method for egomotion estimation from noisy flow fields. Figure best viewed in color. (A) Example optical flow field from two frames of KITTI odometry (sequence 5, images 2358-2359). Note the outliers on the grass in the lower right part of the image and scattered throughout the flow field. (B) We evaluate the flow field under M models with translation parameters sampled uniformly over the unit hemisphere. The residuals for the flow field under three counterfactual models are shown. Each black point indicates the translation direction used. Residuals are scaled to [0,1] for visualization. (C) We estimate the likelihood of each observed residual under each of the models by fitting a Laplacian distribution to each set of residuals. The final confidence weight for each flow vector is estimated as the expected value of the residual likelihood over the set of counterfactual models. Likelihood distributions are shown for the three models above. (D) The weighted flow field is used to make a final estimate of the true egomotion parameters. The black point indicates the translation direction estimated using ERL and the green point indicates ground truth. The unweighted estimate of translation is not visible as it is outside of the image bounds.

- Fast UAV maneuvers require an immediate estimate of the direction of translation (the focus of expansion) in order to compute a time-to-collision map.
- Continuous SfM computations result in better estimates when the incoming frame rate is high and the baseline is very small.

However, estimating camera motion and scene parameters from a single camera (*monocular* egomotion estimation) remains a challenging problem. This problem case arises in many contexts where sensor weight and cost are at a premium, as is the case for lightweight UAVs and consumer

cameras. Situations involving monocular sensors on small platforms pose additional problems: computational resources are often very limited and estimates must be made in real time under unusual viewing conditions (e.g. with a vertically flipped camera, no visible ground plane, and a single pass through a scene). These contexts present many sources of noise. Real-time flow estimation produces unreliable data, and the associated noise is often pervasive and non-Gaussian, which makes estimation difficult and explicit outlier rejection problematic. Furthermore, violations of the assumption of scene rigidity due to independent motion of objects in the scene can lead to valid flow estimates that are outliers nonetheless. Even in the noise-free case, camera motion estimation is plagued with many suboptimal interpretations (illusions) caused by the hilly structure of the cost function. Additionally, forward motion, which is very common in real-world navigation, is known to be particularly hard for monocular visual odometry [4].

We propose an algorithm suitable for the robust estimation of camera egomotion and scene depth from noisy flow in real-world settings with high-frame-rate video, large images, and a large number of noisy optical flow estimates. Our method runs in real-time on a single CPU and can estimate camera motion and scene depth in scenes with noisy optical flow with outliers, making it suitable for integration with filters for real-time navigation and for deployment on light-weight UAVs. The technical contributions of this paper are:

- A novel robust estimator based on the expected residual likelihood (ERL) of flow data that effectively attenuates the influence of outlier flow measurements and runs at 30-40 Hz on a single CPU.
- A novel robust optimization strategy using a lifted kernel that modifies the shape of the objective function to enable joint estimation of weights and model parameters, while enabling good empirical convergence properties.

II. Related work

A. Egomotion/visual odometry

Many approaches to the problem of visual odometry have been proposed. A distinction is commonly made between feature-based methods, which use a sparse set of matching feature points to compute camera motion, and direct methods, which estimate camera motion directly from intensity gradients in the image sequence. Feature-based approaches can again be roughly divided into two types of methods: those estimating camera motion from point correspondences between two frames (*discrete* approaches) and those estimating camera motion and scene structure from the optical flow measurements induced by the motion between the two frames (*continuous* approaches). In practice, point correspondences and optical flow measurements are often obtained using similar descriptor matching strategies. Nonetheless, the discrete and continuous approaches use different problem formulations, which reflect differing assumptions about the size of the baseline between the two camera positions.

The continuous approach is the appropriate choice in situations where the real-world camera motion is slow relative to the sampling frequency of the camera. Our approach is primarily intended for situations in which this is the case, e.g. UAVs equipped with high-frame-rate cameras. Accordingly, we focus our review on continuous, monocular methods. For a more comprehensive discussion, see [5].

B. Continuous, monocular approaches

In the absence of noise, image velocities at 5 or 8 points can be used to give a finite number of candidate solutions for camera motion [6] [7] [8]. With more velocities, there is a unique optimal solution under typical scene conditions [9]. Many methods have been proposed to recover this solution, either by motion parallax [6] [10] [11] [12] or by using the so-called continuous epipolar constraint [5]. The problem is nonlinear and nonconvex, but various linear approximation methods have been proposed to simplify and speed up estimation [13] [14] [15].

Although the problem has a unique optimum, it is characterized by many local minima, which pose difficulties for linear methods [16]. Furthermore, in the presence of noise, many methods are biased and inconsistent in the sense that they do not produce correct estimates in the limit of an unlimited number of image velocity measurements [17]. Many methods also fail under many common viewing conditions or with a limited field of view [18]. Recently, [19] and [20] proposed branch-and-bound methods that estimate translational velocity in real time and effectively handle a large numbers of outliers. However, these methods deal with the case of pure translational camera motion, while our approach estimates both translational and rotational motion.

Most directly related to our work is the robust estimation framework presented in [21]. They propose a method based on a variant of a common algebraic manipulation and show that this manipulation leads to an unbiased, consistent estimator. They pose monocular egomotion as a nonlinear least-squares problem in terms of the translational velocity. In this framework, angular velocity and inverse scene depths are also easily recovered after translational velocity is estimated. To add robustness, they use a loss function with sub-quadratic growth, which they solve by iteratively reweighted least squares (IRLS). We use a similar formulation but demonstrate several novel methods for estimating the parameters of a robust loss formulation. Our methods have properties that are well-suited for dealing with image sequences containing several thousand flow vectors in real time. In particular, we demonstrate that the ERL method adds robustness without requiring costly iterative reweighting, resulting in very little runtime overhead.

Other methods for monocular odometry augment velocity data with planar homography estimates [22] [23] or depth filters [2] to estimate scale. In this work, we do not rely on ground-plane estimation in order to maintain applicability to cases such as UAV navigation, where image sequences do not always contain the ground plane. Because we focus on frame-by-frame motion estimation, we cannot rely on a filtering

Algorithm 1 ERL confidence weight estimation

Input: Measured flow $\{u_n\}_{n=1}^N$, sampled translational velocities $\{t_m\}_{m=1}^M$
Output: Estimated confidence weights $\{\hat{w}_n\}_{n=1}^N$

for all m **do**

Compute scaled residuals:

$$\tilde{r}_u = |A^\perp(t_m)^\top (B\hat{\omega}_m(t_m) - u)|$$

Compute maximum likelihood estimators of residual distribution:

$$\hat{\mu}_m = \text{median}(\tilde{r}_u)$$

$$\hat{b}_m = \frac{1}{N}\sum_{n=1}^{N} \|\tilde{r}_{u_n} - \hat{\mu}_m\|$$

end for
for all n **do**

Compute confidence weights as expected likelihood under Laplacian fits:

$$\hat{w}_n = \frac{1}{M}\sum_{m=1}^{M} \mathscr{L}(\tilde{r}_{u_n}; \hat{\mu}_m, \hat{b}_m)$$

end for
return $\{\hat{w}_n\}_{n=1}^N$

approach to estimate depth. Our method can be augmented with domain-appropriate scale or depth estimators as part of a larger SLAM system.

C. Robust optimization

In this work, we propose to increase the robustness of monocular egomotion estimation (1) by estimating each flow vector's confidence weight as its expected residual likelihood (ERL) and (2) by using a lifted robust kernel to jointly estimate confidence weights and model parameters. ERL confidence weights are conceptually similar to the weights recovered in the IRLS method for optimizing robust kernels [24]. Robust kernel methods attempt to minimize the residuals of observations generated by the target model process ("inliers") while limiting the influence of other observations ("outliers"). Such methods have been used very successfully in many domains of computer vision [25] [26]. However, we are unaware of any previous work that attempts to estimate confidence weights based on the distribution of residuals at counterfactual model parameters, as we do in the ERL method.

The lifted kernel approach offers another method to design and optimize robust kernels in particularly desirable ways. Lifted kernels have recently been used in methods for bundle adjustment in SfM [27], object pose recovery [28], and non-rigid object reconstruction [29]. Our lifted kernel approximates the truncated quadratic loss, which has a long history of use in robust optimization in computer vision [30] and has demonstrated applicability in a wide variety of problem domains.

Previous studies have used robust loss functions for monocular egomotion [21], visual SLAM [31], and RGB-D odometry [32]. To our knowledge, we present the first application of lifted kernels for robust monocular egomotion. Noise is typically handled in odometry by using sampling-based iterative methods such as RANSAC, which makes use of a small number of points to estimate inlier sets (typically five or eight points in monocular methods). The use of a robust kernel allows us to derive our final estimate from a larger number of points. This is desirable because the structure of the problem of continuous monocular odometry admits fewer correct solutions when constrained by a larger number of input points, which can better reflect the complex depth structure of real scenes. Our robust methods allow us to take advantage of a large number of flow estimates, which, while noisy, may each contribute weakly to the final estimate.

III. Problem formulation and approach

In this section, we present the continuous formulation of the problem of monocular visual egomotion. We describe and motivate our approach for solving the problem in the presence of noisy optical flow. We then describe two methods for estimating the confidence weights for each flow vector in a robust formulation of the problem, as well as the pipeline we use to estimate camera motion and scene depth.

A. Visual egomotion computation and the motion field

In the continuous formulation, visual egomotion methods attempt to estimate camera motion and scene parameters from observed local image velocities (optical flow). The velocity of an image point due to camera motion in a rigid scene under perspective projection is given by

$$u(x_i) = \rho(x_i)A(x_i)t + B(x_i)\omega. \tag{1}$$

where $u_i(x_i) = (u_i, v_i)^\top \in \mathbf{R}^2$ is the velocity (optical flow) at image position $x_i = (x_i, y_i)^\top \in \mathbf{R}^2$, $t = (t_x, t_y, t_z)^\top \in \mathbf{R}^3$ is the camera's instantaneous translational velocity, $\omega = (\omega_x, \omega_y, \omega_z)^\top \in \mathbf{R}^3$ is the camera's instantaneous rotational velocity, and $\rho(x_i) = \frac{1}{Z(x_i)} \in \mathbf{R}$ is the inverse of scene depth at x_i along the optical axis. We normalize the camera's focal length to 1, without loss of generality. In the case of calibrated image coordinates,

$$A(x_i) = \begin{bmatrix} 1 & 0 & -x_i \\ 0 & 1 & -y_i \end{bmatrix},$$

$$B(x_i) = \begin{bmatrix} -x_i y_i & 1+x_i^2 & -y_i \\ -1-y_i^2 & x_i y_i & x_i \end{bmatrix}.$$

This formulation is appropriate for the small-baseline case where point correspondences between frames can be treated as 2D motion vectors.

The goal of monocular visual egomotion computation is thus to estimate the six motion parameters of t and ω and the N values for ρ from N point velocities u induced by camera motion. t and ρ are multiplicatively coupled in equation (1) above, so t can only be recovered up to a scale. We therefore restrict estimates of t to the unit hemisphere, $\|t\| = 1$.

The full expression for the set of N point velocities can be expressed compactly as

$$u = A(t)\rho + B\omega. \tag{2}$$

where the expressions for $A(x)$, $B(x)$, and $\rho(x)$ for all N points are

$$A(t) = \begin{bmatrix} A(x_1)t & 0 & \dots & 0 \\ 0 & A(x_2)t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A(x_N)t \end{bmatrix} \in \mathbf{R}^{2N \times N}$$

$$B = \begin{bmatrix} B(x_1) \\ B(x_2) \\ \vdots \\ B(x_N) \end{bmatrix} \in \mathbf{R}^{2N\times 3}$$

and the velocity and depth for each of the points are concatenated to form $u = (u_1^\top, u_2^\top, \ldots, u_N^\top)^\top \in \mathbf{R}^{2N\times 1}$ and $\rho = (\rho(x_1), \rho(x_2), \ldots, \rho(x_N))^\top \in \mathbf{R}^{N\times 1}$. We estimate camera motion and scene depth by minimizing the objective

$$\begin{aligned} \min_{t,\rho,\omega} E(t,\rho,\omega) &= \min_{t,\rho,\omega} L(r(t,\rho,\omega)) \\ &= \min_{t,\rho,\omega} \|A(t)\rho + B\omega - u\|_2^2. \end{aligned} \tag{3}$$

Here, $L(x) : \mathbf{R}^N \to \mathbf{R}$ is a loss function and $r(t,\rho,\omega) : \mathbf{R}^{N+6} \to \mathbf{R}^N$ is a residual function for the flow field depending on the estimated model parameters. We first describe the case of an unweighted residual function under a quadratic loss, which is suitable for the case of Gaussian noise.

Following [21], we note that no loss of generality occurs by first solving this objective for ρ in the least-squares sense. Minimizing over ρ gives

$$\begin{aligned} &\min_{t,\omega} \min_{\rho} \|A(t)\rho + B\omega - u\|_2^2 \\ &= \min_{t,\omega} \|A^\perp(t)^\top (B\omega - u)\|_2^2, \end{aligned} \tag{4}$$

where $A^\perp(t)$ is the orthogonal compliment to $A(t)$. This expression no longer depends on ρ and depends on t only through $A^\perp(t)^\top$, which is fast to compute due to the sparsity of $A(t)$ (see section II of the supplement for more details).

In the absence of noise, we could proceed by directly minimizing equation (4) in t and ω. In particular, given a solution for t, we can directly solve for ω by least squares in $O(N)$ time. In the noiseless case, we estimate t by optimizing

$$\min_t \|A^\perp(t)^\top (B\hat{\omega}(t) - u)\|_2^2, \tag{5}$$

where $\hat{\omega}(t)$ is the least-squares estimate of ω for a given t (see section IV of the supplement for more details). This method of estimating t, ρ, and ω was shown to be consistent in [17]. That is, in the absence of outliers, this method leads to arbitrarily precise, unbiased estimates of the motion parameters as the sample size increases.

B. Robust formulation

However, the manipulations introduced in equations (4) and (5) rely on least-squares solutions and are not stable in the presence of outliers. Accordingly, instead of directly solving (5), we propose to solve a robust form. To do so, we introduce a confidence weight for each flow vector $w_i(u_i) \in [0,1]$ to give

$$\begin{aligned} &\min_t L(r(t,\hat{\omega}(t)), w) \\ &= \min_t \|w \circ A^\perp(t)^\top (B\hat{\omega}(t) - u)\|_2^2, \end{aligned} \tag{6}$$

where $w = (w(u_1), w(u_2), \ldots, w(u_N))^\top \in [0,1]^N$ is the vector of all weights, $r \in \mathbf{R}^N$ is the vector of residuals for the flow field at some estimate of t, and $\circ$ is the Hadamard product.

Each entry $w(u_i)$ of w attempts to weight the corresponding data point u_i proportionally to its residual at the optimal model parameters $(\hat{t}, \hat{\rho}, \hat{\omega})$, reflecting the degree to which the point is consistent with a single generating function for the motion in the scene, possibly with Gaussian noise. In other words, it reflects the degree to which u_i is an inlier for the optimal model of camera motion in a rigid scene. This is equivalent to replacing the choice of $L(x) = x^2$ as the loss in equation (5) with a function that grows more slowly.

We introduce a method to directly estimate the confidence weights as the expected residual likelihood (ERL) for each flow vector given the distribution of residuals for the flow field at a range of model parameters consistent with the solution in (5). We interpret each weight in terms of an estimate of the validity of the corresponding point under the model: that is, as an estimate of the point's residual at the optimal model parameters in a noise-free context. We compare ERL to a method that replaces $L(x) = x^2$ in (5) with a lifted truncated quadratic kernel [27] and jointly optimizes the confidence weights and model parameters. We demonstrate that ERL outperforms the lifted kernel approach on the KITTI dataset, and both of these approaches outperform existing methods for monocular egomotion computation.

C. Confidence weight estimation by expected residual likelihood

Here, we describe the ERL method for estimating the confidence weights in (6), and we demonstrate that this method provides a good estimate of the appropriate confidence weights in the case of optical flow for visual egomotion.

At the optimal model parameters, (t^*, ρ^*, ω^*), the residuals for inlier points (i.e. correct flow vectors due to rigid motion) are distributed according to a normal distribution, reflecting zero-mean Gaussian noise. However, in the presence of outliers, a zero-mean Laplacian distribution provides a better description of the residual distribution (see **Supplemental Fig. 2**). Accordingly, we can fit a Laplacian distribution to the observed residuals at the optimal model parameters to approximate the probability density function for residuals.

We use this property to identify outliers as those points that are inconsistent with the expected residual distribution at a range of model values. For each point, we compute the likelihood of each observed, scaled residual as

$$p(\tilde{r}_{u_i}^m | (t_m, \rho_m, \omega_m), \tilde{r}_u^m) = \mathscr{L}(\tilde{r}_{u_i}; \hat{\mu}_m, \hat{b}_m), \tag{7}$$

where $\tilde{r}_{u_i}^m$ is the scaled residual under the m$^{\text{th}}$ model (t_m, ρ_m, ω_m) at the i$^{\text{th}}$ flow vector and $\tilde{r}_u^m = (\tilde{r}_{u_1}^m, \tilde{r}_{u_2}^m, ..., \tilde{r}_{u_N}^m)^\top$. We fit $\hat{\mu}_m$ and $\hat{b}_m$, respectively the location and scale parameters of the Laplacian distribution, to the set of scaled residuals $\tilde{r}_u^m$ using maximum likelihood.

Because inliers exhibit smaller self-influence than outliers [33], inlier residuals will typically be associated with higher likelihood values. However, the distribution used to estimate the likelihood reflects both the inlier and outlier points. If the counterfactual model parameters used to estimate the m$^{\text{th}}$ likelihood correspond to a model that is highly suboptimal, some outliers may be assigned higher likelihoods than they

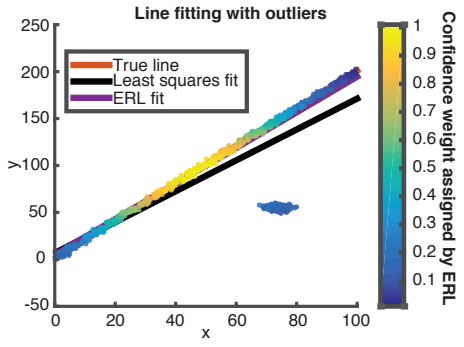

Fig. 2. A 2D line-fitting problem demonstrating how ERL weights inliers and outliers. Inliers are generated as $y_i \approx 2x_i + 1$ with Gaussian noise. Each data points is colored according to its estimated confidence weight.

would be at the optimal model. Moreover, the presence of Gaussian noise means that the estimated likelihood for individual inliers may be erroneously low by chance for a particular model even if the optimal exponential distribution is exactly recovered.

To arrive at more reliable estimates and to discount the effect of erroneous likelihoods due to the specific model parameters being evaluated, we estimate the expected residual likelihood for each data point by evaluating the likelihood under M models,

$$\hat{w}_i = \mathbb{E}[\tilde{r}_{u_i}^m] = \frac{1}{M}\sum_{m=1}^{M} \mathscr{L}(\tilde{r}_{u_i}; \hat{\mu}_m, \hat{b}_m). \tag{8}$$

This method returns a vector $\hat{w} \in \mathbf{R}^N$. To use $\hat{w}$ as confidence weights in a robust optimization context, we scale them to the interval $[0,1]$. Scaling the maximum $\hat{w}_i$ to 1 and the minimum $\hat{w}_i$ to 0 for each flow field works well in practice.

The full process to estimate weights by ERL is shown in **Algorithm 1**. This method returns confidence weights in $O(MN)$ time, where M is set by the user. Empirically, the ERL method gives results that reflect the inlier structure of the data with small values of M (we use $M \approx 100$), allowing very quick runtimes. In practice, the method assigns high weights to very few outliers while assigning low weights to acceptably few inliers. Thus, the method balances a low false positive rate against a moderately low false negative rate. This is a good strategy because our method takes a large number of flow vectors as input, which leads to redundancy in the local velocity information. **Fig. 2** illustrates the ERL method's use in a simple 2D robust line-fitting application.

As discussed above, choosing values for the confidence weights in a least squares objective is equivalent to fitting a robust kernel. We note that regression under the assumption of Laplacian noise leads to an L1 cost. However, we have no guarantees about the form of the robust kernel corresponding to the weights chosen by the ERL method. Accordingly, we also explored using a robust kernel with known properties.

D. Robust estimation using a lifted kernel

Here, we explore the effect of jointly optimizing the confidence weights, $w(u)$, and ω for a given value of t using

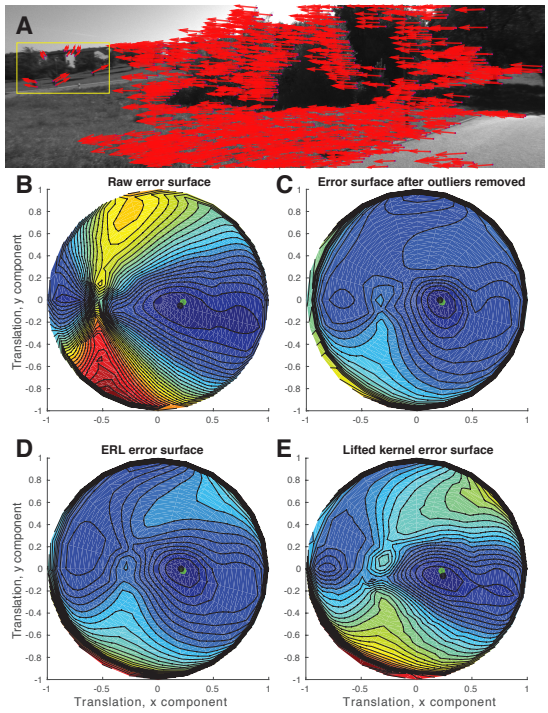

Fig. 3. Robust methods recover the error surface of the outlier-free flow field. (A) Example optical flow field from two frames of KITTI odometry (sequence 10, images 14-15). Note the prominent outliers indicated by the yellow box. Error surfaces on this flow field for (A) the raw method (equation (5)) with all flow vectors, (B) with outliers removed by hand, and (C) with confidence weights estimated by ERL or (D) the lifted kernel. The green point is the true translational velocity and the black point the method's estimate. Blue: low error. Red: high error. Translation components are given in calibrated coordinates.

the lifted kernel approach described in [27]. In our case, a lifted kernel takes the form

$$\min_{t,\omega,w} \hat{L}(r(t,\omega),w)$$

$$= \min_{t} \min_{\omega,w} (\|w \circ A^{\perp}(t)^{\top}(B\omega(t) - u)\|_2^2 + \sum_{i=1}^{N} \kappa^2(w_i^2)), \tag{9}$$

where the lifted kernel of the loss L is denoted as $\hat{L}$. $\kappa(x)$: $\mathbf{R} \to \mathbf{R}$ is a regularization function applied to the weights. Because this approach does not rely on the least squares solution for rotational velocity, $\hat{\omega}$, it may gain additional robustness to noise. This approach also allows us to estimate the confidence weights for particular values of t, unlike the ERL approach, which relies on estimates at several values of t to produce stable results.

Different choices of κ produces different kernels. We use

$$\kappa(w^2) = \frac{\tau}{\sqrt{2}}(w^2 - 1), \tag{10}$$

which gives a kernel that is a smooth approximation to the truncated quadratic loss [27]. τ is a hyperparameter that determines the extent of the quadratic region of the truncated quadratic loss. We set $\tau = 0.05$ for all results shown here, but other choices give similar results.

The lifted kernel approach to solving nonlinear least squares problems is similar to IRLS insofar as it incorporates

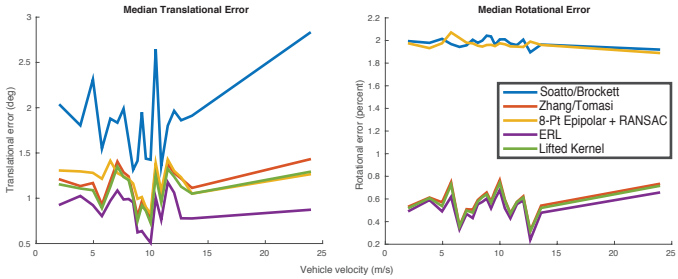

Fig. 4. Median translational and rotational errors on the full KITTI odometry dataset for our methods and baselines.

confidence weights on each of the data points and optimizes the values of these weights in addition to the value of the target model parameters. However, rather than alternately estimating the best weights given estimated model parameters and the best model parameters given estimated weights, the lifted approach simultaneously optimizes for both weights and model parameters, effectively "lifting" a minimization problem to a higher dimension.

The lifted kernel approach has several properties that are particularly beneficial for encouraging fast convergence. First, by using the weights to increase the dimensionality of the optimization problem, the lifted kernel minimizes the extent of regions of low gradient in the cost function. This ensures the method can quickly and reliable converge to minima of the function. Second, optimization can exploit the Gauss-Newton structure of the joint nonlinear least-squares formulation for faster convergence than the slower iterative-closest-points-like convergence exhibited by IRLS.

To illustrate the effect of our two robust optimization strategies, we display the error surfaces for the ERL and lifted-kernel methods on a sample flow field from KITTI (**Fig. 3**). The error surfaces are shown as a function of the translational velocity. Both methods recover error surfaces that resemble the error due to inlier flow vectors. The confidence weights estimated by ERL generally more closely resemble the pattern of inliers and outliers in flow data. To produce the results for the case with outliers removed, we strengthened the maximum bidirectional error criterion for flow inclusion to eliminate noisy matches and manually removed obvious outliers from the flow field.

IV. Experiments

We compare the performance of the proposed methods (called "ERL" and "Lifted Kernel" in the figures) to several baseline methods for monocular egomotion/visual odometry from the literature: 5-point epipolar+RANSAC (using [34]), 8-point epipolar+RANSAC (using [35]), and two continuous epipolar methods - Zhang/Tomasi [21], which is identical to equation (5), and Soatto/Brockett [36]. All experiments were run on a desktop with an Intel Core i7 processor and 16 GB of RAM. A single CPU core was used for all experiments.

With $\sim$1000 flow vectors, the ERL method runs at 30-40 Hz in an unoptimized C++ implementation. Because of the low overhead of the ERL procedure, this is effectively the same runtime as the Zhang/Tomasi method. The lifted kernel optimization has no convergence guarantees, and it typically runs at $<$1 Hz in a MATLAB implementation. Note that both of these runtimes can be significantly improved with better optimization. The Soatto/Brockett method runs extremely quickly ($>$500 Hz), but performs poorly on real sequences. The implementation of epipolar+RANSAC used here runs at $\sim$25 Hz. Optical flow for all our results was extracted using a multiscale implementation of the KLT method [37] [38].

For both ERL and the lifted approach, we optimize t using Gauss-Newton. We initialize t at a grid of values spaced over the unit hemisphere to decrease the chance of converging to a non-global minimum. We then prune the grid to a single initial value t_0 by choosing the grid point that gives the lowest residual under equation (6) or (9) for ERL or the lifted kernel, respectively. We then optimize to convergence starting from t_0. This pruning strategy is effective at avoiding local minima because good estimates for the weights return an error surface that is very similar to the noiseless case (see **Fig. 3**) and this error surface is smooth with respect to the sampling density we use (625 points) [16]. Confidence weights for ERL are computed using model parameters sampled on a coarser grid (100 points), as this is adequate to give good confidence weight estimates.

For all tests using the lifted kernel, we optimize the expression in equation (9) using the efficient Schur compliment implementation of Levenberg-Marquardt described in [27]. Details of the optimization procedure used here are given in section III of the supplement. We did not explore jointly optimizing over t, ω, and w, but joint optimization over these model parameters with a lifted kernel is possible, and we plan to explore its use in future work.

A. Evaluation on KITTI

We evaluate the performance of our method using the KITTI dataset [39], which is a collection of real-world driving sequences with ground-truth camera motion and depth data. The sequences contained in the dataset are challenging for state-of-the-art odometry methods for several reasons. First, they contain large inter-frame motions and repetitive scene structures that make estimating accurate flow correspondences difficult in real time. Second, several sequences feature little to no camera motion, which typically causes monocular odometry methods to fail. Finally, some sequences contain independent motion due to other vehicles and pedestrians, which violates the assumption of scene rigidity and makes reliable odometry more difficult.

All results are performed on neighboring frames of the KITTI odometry dataset (no skipped-frame sequences are evaluated), as these image pairs better match the modeling assumptions of continuous egomotion/odometry methods. All sequences were captured at 10 Hz at a resolution of 1392 x 512 pixels. We evaluated all methods on all 16 sequences of the KITTI odometry test set.

The results for methods on KITTI are shown in **Figs. 4-6**. For ease of visualization, the results for the 5-point

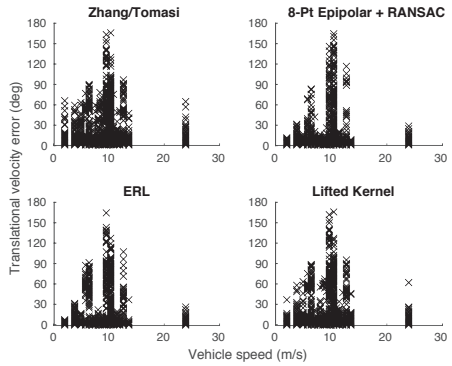

Fig. 5. Full distribution of translational velocity errors.

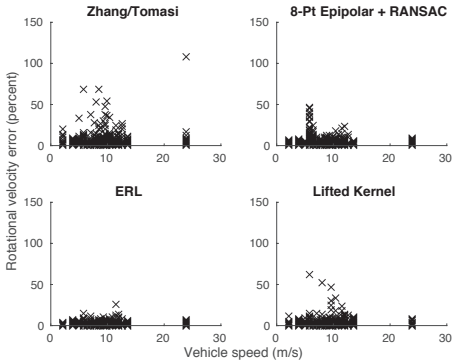

Fig. 6. Full distribution of rotational velocity errors.

epipolar method with RANSAC are not shown (they were significantly worse than all other methods we attempted). ERL produces the best estimates of translational velocity, while the lifted kernel produces results of similar quality to 8-point epipolar with RANSAC and the Zhang/Tomasi method. ERL, the lifted kernel, and Zhang/Tomasi produce rotational velocity estimates of similar quality. The 8-point epipolar method produces worse estimates in this case because of the large baseline assumption, which is not suitable for rotational velocity estimation under these conditions. Soatto/Brockett produces bad estimates in these test cases because of the bias introduced by its algebraic manipulation.

B. Synthetic sequences

To estimate the robustness of our methods to outliers, we test the methods on synthetic data. Synthetic data were created by simulating a field of 1500 image points distributed uniformly at random depths between 2 and 10 m in front of the camera and uniformly in x and y throughout the frame. A simulated camera is moved through this field with a translational velocity drawn from a zero-mean Gaussian with standard deviation of 1 m/frame and a rotational velocity drawn from a zero-mean Gaussian with standard deviation of 0.2 radians/frame. Flow was generated from the resulting 3D point trajectories by perspective projection using a camera model with a 1 m focal length. All flow vectors were corrupted with noise in a random direction and magnitude drawn from a zero-mean Gaussian with a standard deviation $1/10^{\text{th}}$ the mean flow vector magnitude. Outliers were created by replacing a fraction of the points with random values drawn from a Gaussian fit to the magnitude and direction of all inlier flow vectors. We ran 100 iterations at each outlier rate. We ran all egomotion methods on the same data.

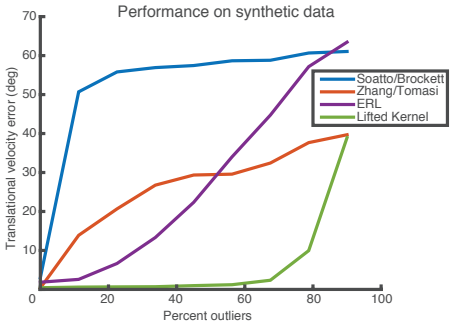

Fig. 7. Translation error as a function of percent outliers on synthetic data for our robust methods and two baseline continuous egomotion methods.

The errors in translational motion estimated on this data are shown in **Fig. 7**. As expected, the two robust methods outperform least-squares methods for reasonable numbers of outliers. At higher outlier rates, however, the performance of both robust methods deteriorates. Interestingly, the performance of the lifted kernel method is stable even when the majority of data points are outliers. We are uncertain why the lifted kernel performs better than ERL on synthetic data, while the opposite is true for KITTI. This difference may be due to the way the data were generated - in KITTI, outliers often reflect real structures in the scene and may contain some information about camera motion, but this is not the case in the synthetic data. The difference may also be due in part to the difference in depth structures in KITTI and the synthetic data. In KITTI, flow magnitude for both inliers and outliers is reflective of depth structure, and depth in real scenes is not distributed uniformly.

V. Conclusions

We have introduced new techniques for robust, continuous egomotion computation from monocular image sequences. We described ERL, a novel robust method that directly estimates confidence weights for the vectors of a flow field by evaluating the distribution of flow residuals under a set of self-consistent counterfactual model parameters. We also introduced a new formulation of the perspective motion equation using a lifted kernel for joint optimization of model parameters and confidence weights. We compared the results of ERL and the lifted kernel formulation, and showed that while the lifted kernel appears to be more stable in the presence of a large fraction of outliers, ERL performs better in a real-world setting. The ERL method achieves good results on KITTI without relying on stereo data or ground plane estimation and accordingly is well-suited for use in lightweight UAV navigation. We are unable to directly evaluate our methods on this target domain because there are currently no UAV datasets with suitable ground truth.

Although the empirical results here are promising, we have no guarantees on the weights recovered by ERL, and this remains a topic for future work.

Our code is publicly available at `https://github.com/stephenphillips42/erl_egomotion`.

Acknowledgments

The authors gratefully acknowledge support by the grants NSF-DGE-0966142, NSF-IIP-1439681, NSF-IIS-1426840, ARL MAST-CTA W911NF-08-2-0004, and ARL RCTA W911NF-10-2-0016.

References

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *Robotics & Automation Magazine, IEEE*, vol. 18, no. 4, pp. 80–92, 2011.

[2] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 15–22.

[3] P.-J. Bristeau, F. Callou, D. Vissiere, N. Petit, *et al.*, "The navigation and control technology inside the ar.drone micro uav," in *18th IFAC world congress*, vol. 18, no. 1, 2011, pp. 1477–1484.

[4] J. Oliensis, "The Least-Squares Error for Structure from Infinitesimal Motion," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 259–299, Feb. 2005.

[5] Y. Ma, *An invitation to 3-d vision: from images to geometric models*. springer, 2004, vol. 26.

[6] H. C. Longuet-Higgins and K. Prazdny, "The interpretation of a moving retinal image," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 208, no. 1173, pp. 385–397, 1980.

[7] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[8] D. Nistér, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 756–770, 2004.

[9] B. K. P. Horn, "Motion fields are hardly ever ambiguous," *International Journal of Computer Vision*, vol. 1, no. 3, pp. 259–274, Oct. 1988.

[10] E. C. Hildreth, "Recovering heading for visually-guided navigation," *Vision Research*, vol. 32, no. 6, pp. 1177–1192, June 1992.

[11] D. J. Heeger and A. D. Jepson, "Subspace methods for recovering rigid motion I: Algorithm and implementation," *International Journal of Computer Vision*, vol. 7, no. 2, pp. 95–117, 1992.

[12] A. D. Jepson and D. J. Heeger, "Subspace methods for recovering rigid motion II: Theory," Technical Report RBCV-TR-90-36, University of Toronto, Tech. Rep., 1990.

[13] A. Jepson and D. Heeger, "A fast subspace algorithm for recovering rigid motion," in *, Proceedings of the IEEE Workshop on Visual Motion, 1991*, 1991, pp. 124–131.

[14] X. Zhuang, T. S. Huang, N. Ahuja, and R. M. Haralick, "A simplified linear optic flow-motion algorithm," *Computer Vision, Graphics, and Image Processing*, vol. 42, no. 3, pp. 334–344, 1988.

[15] K. Kanatani, "3-D interpretation of optical flow by renormalization," *International Journal of Computer Vision*, vol. 11, no. 3, pp. 267–282, 1993.

[16] A. Chiuso, R. Brockett, and S. Soatto, "Optimal Structure from Motion: Local Ambiguities and Global Estimates," *International Journal of Computer Vision*, vol. 39, no. 3, pp. 195–228, Sept. 2000.

[17] T. Zhang and C. Tomasi, "On the consistency of instantaneous rigid motion estimation," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 51–79, 2002.

[18] K. Daniilidis and H.-H. Nagel, "Analytical results on error sensitivity of motion estimation from two views," *Image and Vision Computing*, vol. 8, no. 4, pp. 297–303, 1990.

[19] J. Fredriksson, O. Enqvist, and F. Kahl, "Fast and Reliable Two-View Translation Estimation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1606–1612.

[20] J. Fredriksson, V. Larsson, and C. Olsson, "Practical Robust Two-View Translation Estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2684–2690.

[21] T. Zhang and C. Tomasi, "Fast, robust, and consistent camera motion estimation," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[22] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, June 2011, pp. 963–968.

[23] S. Song, M. Chandraker, and C. Guest, "Parallel, real-time monocular visual odometry," in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, May 2013, pp. 4698–4705.

[24] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics - Theory and Methods*, vol. 6, no. 9, pp. 813–827, Jan. 1977.

[25] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 367–383, 1992.

[26] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *International Journal of Computer Vision*, vol. 19, no. 1, pp. 57–91, 1996.

[27] C. Zach, "Robust Bundle Adjustment Revisited," in *Computer Vision ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, Sept. 2014, no. 8693, pp. 772–787.

[28] C. Zach, A. Penate-Sanchez, and M.-T. Pham, "A Dynamic Programming Approach for Fast and Robust Object Pose Recognition from Range Images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 196–203.

[29] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and others, "Real-time Non-rigid Reconstruction using an RGB-D Camera," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 156, 2014.

[30] A. Blake and A. Zisserman, *Visual reconstruction*. MIT press Cambridge, 1987, vol. 2.

[31] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2320–2327.

[32] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3748–3754.

[33] P. J. Huber, *Robust statistics*. Springer, 2011.

[34] H. Stewenius, C. Engels, and D. Nistér, "Recent developments on direct relative orientation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 60, no. 4, pp. 284–294, 2006.

[35] P. I. Corke, *Robotics, Vision & Control: Fundamental Algorithms in Matlab*. Springer, 2011.

[36] S. Soatto and R. Brockett, "Optimal structure from motion: Local ambiguities and global estimates," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998, pp. 282–288.

[37] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.

[38] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ., 1991.

[39] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.