Multi-objective Visual Odometry

Hsiang-Jen Chien[1(✉)], Jr-Jiun Lin[2], Tang-Kai Yin[2], and Reinhard Klette[1]

[1] Department of Electrical and Electronic Engineering,
School of Engineering, Computer, and Mathematical Sciences,
Auckland University of Technology, Auckland, New Zealand
jchien@aut.ac.nz

[2] Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung, Taiwan

Abstract. Visual odometry (VO) has been extensively studied in the last decade. Despite a variety of implementation details, the proposed approaches share the same principle - a minimisation of a carefully chosen energy function. In this paper we review four commonly adopted energy models including perspective, epipolar, rigid, and photometric alignments, and propose a novel VO technique that unifies multiple objectives for outlier rejection and egomotion estimation to outperform mono-objective egomotion estimation. The experiments show an improvement above 50% is achievable by trading off 15% additional computational cost.

1 Introduction

Visual odometry (VO) uses an image sequence for calculating continuously egomotion of the camera. VO has been actively studied in the fields of computer vision, photogrammetry, or robotics. Egomotion estimation can be approached in a variety of ways. When dense depth data is available (e.g. from a ToF camera), the inter-frame pose can be derived by means of the alignment of 3D-to-3D structure correspondences. If the sensor also provides intensity images (e.g. an RGB-D camera), the pose can be estimated by minimising the photometric error when applied to perspectively warping the images. It is a more general case where 3D coordinates of sparse pixels are known only in the previous frame, where their locations need to be tracked in the next frame. Given such 3D-to-2D correspondences, egomotion is estimated by minimisation of the geodesic reprojection error.

In this paper we provide a review on adopted energy models of state-of-the-art VO methods. Based on these models, we propose a novel VO implementation that collaboratively uses multiple energy models to achieve more robust and accurate egomotion estimation. The rest of this paper is organised as follows. Section 2 gives a brief review on the recent development of VO techniques. In Sect. 3 we formulate VO as an energy minimisation problem. Section 4 walks through the energy models used by the state-of-the-art methods, based on which a unified framework is proposed in Sect. 5. Experimental results and discussions are given in Sect. 6, while Sect. 7 concludes this paper.

M. Paul et al. (Eds.): PSIVT 2017, LNCS 10749, pp. 62–74, 2018.
https://doi.org/10.1007/978-3-319-75786-5_6

2 Literature Review

In the last two decades, the development of VO has led to a separation into two different paths, namely appearance-based or feature-based techniques [1].

Appearance-based VO makes direct use of image intensities to minimise the photometric error between the perspective warping of a referenced frame and the image of a target frame. Early direct methods are influenced by optical flow estimation or *structure-from-motion* (SfM) techniques from the photogrammetry community [3,4]. After decades of oblivion, the direct methods have quickly become popular in the last few years, thanks to the advance of GPU computing and breakthroughs in the sparse *visual simultaneous localisation and mapping* (V-SLAM) domain [5,7–11].

In 2007 the first symbolic implementation of sparse direct VO is presented in the context of augmented reality [5], based on work published in 2001 [4]. In 2011 Davison et al. demonstrated a regularised photometric error function using the gradients of depth maps and intensity images to achieve accurate dense matching over multiple short-baseline movements [7]. For finding the optimal motion that minimises the regularised matching cost, the authors proposed a forward-compositional cost function.

A similar inverse compositional formulation is used in [8] during the iterative Gaussian minimisation over a photometric energy function derived from a number of 4×4 perspectively warped patches around the tracked key points. In [9] the error covariance is taken into account to build normalised photometric error terms. The uncertainty of each tracked key point is propagated by a Jacobian-based approximation, and continuously maintained following an update model that makes joint use of a dense depth map and its variance map. In the authors' follow-up work [10,11], the image gradient as well as the camera's photometric calibration are taken into account to revise the uncertainty estimation model for the photometric energy function.

Feature-based techniques, on the other side, keep tracking a set of distinctive scene points and derive the camera's egomotion from the correspondences established by descriptor matching. Approaches in this category have been dominating the development of VO since its early success on Mars [12]. The implementation on the Mars rovers uses a weighted 3D rigid point alignment model to optimise the estimation of rover's egomotion. In difference to the direct methods, feature correspondences are established in feature space.

In [13], robust outlier rejection is used to remove noisy correspondences. In [14], the observations of each tracked feature are integrated over time to yield more accurate state estimation. The technique is later generalised by [15]. A more recent work [16] demonstrates the feasibility of real-time feature extraction, matching, and pose estimation using oriented FAST key points and rotated BRIEF descriptors. The success of all these methods lies in the minimisation of the geodesic distances between observed feature locations and their predictions.

Few recent work attempts to fill in the gap between appearance-based and feature-based VO or V-SLAM. For example, Forster et al. deployed a reprojection minimisation technique, which is commonly used in feature-based VO, to refine the pose estimated from direct photometric alignment [8]. Another example is the bi-objective energy model used in [10] that takes into account not only a photometric error but also geodesic displacement. Such a trend has inspired this work to study the use of multiple objectives for egomotion estimation.

3 Visual Odometry

Historically, the estimation of a camera's egomotion relies on the tracking of some long-term key points, which are projections of distinctive scene points known as *landmarks*, and the minimisation of the deviation between their predicted locations and actual observations.

3.1 Theory

Let $P = (X, Y, Z)$ be the 3D coordinates of a landmark. Following the pinhole camera model, its projection (x, y) in the image plane is given by

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \sim \begin{pmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = (\mathbf{K}\,\mathbf{0}) \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{1}$$

where the upper triangular matrix $\mathbf{K}$ is the *camera matrix* modelled by the intrinsic parameters of the camera including focal lengths f_x and f_y, and the image centre or principal point (c_x, c_y). By $\sim$ we denote projective equality (i.e. equality up to a scale).

As the camera moves, a new coordinate system is instantiated. The egomotion of the camera can then be modelled by a Euclidean transformation $\mathbf{T} \in \mathbb{SE}(3)$, from the previous frame to the new coordinate system, which consists of a rotation $\mathbf{R} \in \mathbb{SO}(3)$ and a translation component $\mathbf{t} \in \mathbb{R}^3$. If the landmark P remains stationary, its projection into the new camera position can be predicted by

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} \sim \mathbf{K} \left(\mathbf{R}\,\mathbf{t}\right) \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{2}$$

The estimation of an unknown transformation $(\mathbf{R}\,\mathbf{t})$, given a set of 3D-to-2D correspondences $(X, Y, Z) \leftrightarrow (x', y')$, is known as the *perspective-from-n-points* (PnP) problem [18]. Such a problem has been extensively studied in the context of SfM or VO.

The implementation of visual odometry, however, is not limited to the use of 3D-to-2D point-to-point correspondences. For example, when dense depth data is available, one may alternatively use 3D-to-3D correspondences and replace Eq. (2) by a rigid alignment objective (see Sect. 4) to model the error of a motion hypothesis. In the monocular case, on the other hand, due to a lack of 3D data, a set of 2D-to-2D epipolar constraints is commonly used as the objective.

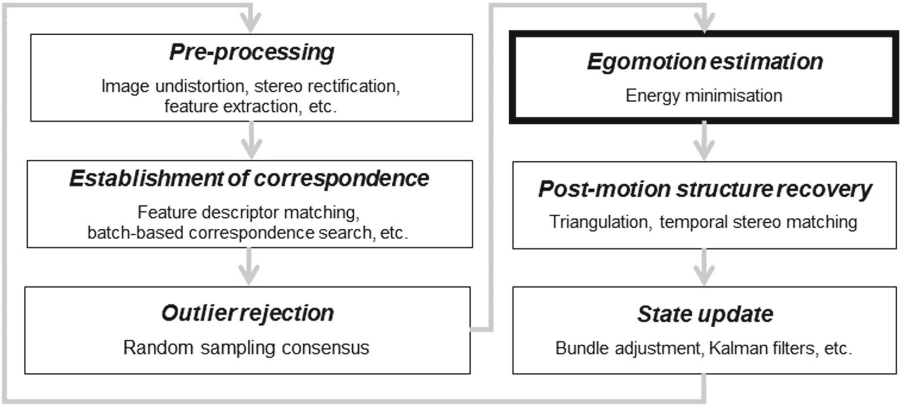
Pre-processing
Image undistortion, stereo rectification, feature extraction, etc.
Establishment of correspondence
Feature descriptor matching, batch-based correspondence search, etc.
Outlier rejection
Random sampling consensus
Egomotion estimation
Energy minimisation
Post-motion structure recovery
Triangulation, temporal stereo matching
State update
Bundle adjustment, Kalman filters, etc.

Fig. 1. General visual odometry model, with each stage annotated by related topics

Figure 1 shows the stages of a generalised visual odometry methodology in an abstract way, independent of the type of correspondences used. In such a general model, the egomotion estimation stage can be conceptualised as a general energy minimisation process.

3.2 Energy Minimisation Problem

As an energy minimisation problem, a residual function $\varphi(\mathbf{x}, \mathbf{y}; \xi) \in \mathbb{R}$ is defined for each established correspondence $\mathbf{x} \leftrightarrow \mathbf{y}$ to solve for egomotion. Note that $\mathbf{x}$ and $\mathbf{y}$ can be any entities of interest, and the residual is parametrised by twist coordinates $\xi \in \mathbb{R}^6$ which is the Lie-algebra entity minimally representing a Euclidean transform $\mathbf{T} = \exp_{\mathrm{se}(3)}(\xi) \in \mathbb{SE}(3)$. Note that two twists ξ and ξ' can be composited by the multiplication of their corresponding Euclidean transformation matrices

$$\xi \circ \xi' = \log_{\mathbb{SE}(3)} \Big(\exp_{\mathrm{se}(3)}(\xi') \cdot \exp_{\mathrm{se}(3)}(\xi) \Big) \tag{3}$$

where $\circ$ is the pose concatenation operator.

Individual residuals are further summarised as a scalar to be minimised. This is often done in the sum-of-squares form to achieve *maximum-likelihood estimation* (MLE) when the error distribution of residuals is believed to follow a Gaussian. Let $\varPhi(\xi) = (\varphi_0, \varphi_1, ..., \varphi_{m-1})$ be an m-vector function instantiated from m correspondences; the optimal estimate of egomotion is found to be

$$\xi = \underset{\hat{\xi} \in \mathbb{R}^6}{\arg\min} (\| \varPhi(\hat{\xi}) \|^2_{\mathbf{\Sigma}}) \tag{4}$$

where $\| \cdot \|^2_{\mathbf{\Sigma}}$ is the squared Mahalanobis distance defined by $\mathbf{\Sigma}$, an $m \times m$ positive definite matrix denoting the error covariance over all the correspondences.

When the correspondences are believed to be established independently (as in most of the cases), $\mathbf{\Sigma}$ is simplified as a diagonal matrix. Equation (4) can then be rewritten as

$$\xi = \underset{\hat{\xi} \in \mathbb{R}^6}{\arg\min} (\sum_i w_i \left\| \varphi_i \left(\mathbf{x}_i, \mathbf{y}_i; \xi \right) \right\|^2) \tag{5}$$

where w_i is the inverse of the i-th diagonal entry in $\mathbf{\Sigma}$. An optimal estimate ξ, that minimises the weighted sum-of-squares, can be approached iteratively by

$$\xi_{k+1} = \varDelta\xi_k \circ \xi_k \tag{6}$$

with the update computed using the Levenberg-Marquardt algorithm [19]:

$$\varDelta\xi_k = \left(\mathbf{H} + \lambda \ \mathrm{diag}(\mathbf{H}) \right)^{-1} \mathbf{J}^\top \varPhi(\xi_k) \tag{7}$$

where $\lambda \in \mathbb{R}$ is the damping variable, and $\mathbf{H} = \mathbf{J}^\top \mathbf{W} \mathbf{J}$ is the Hessian matrix approximated by the weight matrix $\mathbf{W} = \mathrm{diag}(w_0, w_1, ..., w_{i-1})$ and the Jacobian

$$\mathbf{J}_{ij} = \frac{\partial \varphi_i}{\partial \xi_j}(\xi_k). \tag{8}$$

of $\varPhi$ at ξ_k. The variable λ is adaptively adjusted to control the optimisation toward a Gauss-Newton-like process (when ξ is far from a local minimum), or a gradient-descent-like process (when ξ is closer to a local minimum.) All the energy functions considered in this work are minimised in this manner, with the Jacobian matrix numerically computed by first-order finite differentiations.

4 Energy Models in Visual Odometry

In this section we review three geodesic models and one photometric energy model, widely chosen in the literature.

4.1 Epipolar Alignment

Given an image point (x, y) in the current frame and $(\mathbf{R}, \mathbf{t})$ for the motion of the camera, the corresponding epipolar line can be obtained in the next frame identifying the search domain for the corresponding image point (x', y'). Such a 2D-to-2D point correspondence $(x, y) \leftrightarrow (x', y')$ is useful for evaluating the correctness of a motion hypothesis.

Let $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ be a hypothesis, and $(x, y) \leftrightarrow (x', y')$ be the projections of a static scene point. The back-projected rays through these image points have to be co-planar, leading to the well-known epipolar constraint

$$\mathbf{x}'^\top \mathbf{K}^{-\top} [\hat{\mathbf{t}}]_\times \hat{\mathbf{R}} \mathbf{K}^{-1} \mathbf{x} = 0 \tag{9}$$

where $[\hat{\mathbf{t}}]_\times$ is the skew-symmetric matrix form of vector $\mathbf{t}$, $\mathbf{K}$ is the camera matrix, and $\mathbf{x} = (x, y, 1)^\top$ are homogeneous coordinates of an image point in vector form.

In practice, the equality of Eq. (9) never holds, as a result of numerical computations, errors in correspondences, and the inaccuracy of the motion hypothesis. For the last factor, from a set of correspondences $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ one may obtain the residual terms

$$\varphi_i(\mathbf{x}_i, \mathbf{x}'_i; \hat{\xi}) = \mathbf{x}'^{\top}_i \hat{\mathbf{F}} \mathbf{x}_i \tag{10}$$

where $\hat{\mathbf{F}} = \mathbf{K}^{-\top} [\hat{\mathbf{t}}]_{\times} \hat{\mathbf{R}} \mathbf{K}^{-1}$ is the fundamental matrix encoding the given epipolar geometry, and $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ is the motion hypothesis converted from $\hat{\xi}$.

The algebraic distances, however, are found biased as the image points far away from the epipole tend to be over-penalised. A geometrically meaningful modelling is to measure the shortest distance

$$\delta(\mathbf{x}', \mathbf{l}) = \frac{|\mathbf{x}'^{\top} \mathbf{F} \mathbf{x}|}{\sqrt{l_0^2 + l_1^2}} \tag{11}$$

between $\mathbf{x}'$ and the corresponding epipolar line $\mathbf{l} = \mathbf{F}\mathbf{x} = (l_0, l_1, l_2)^{\top}$, for a general correspondence $\mathbf{x} \leftrightarrow \mathbf{x}'$ and fundamental matrix $\mathbf{F}$. As the observation $\mathbf{x}'$ also introduces an epipolar constraint on $\mathbf{x}$, we have that

$$\delta(\mathbf{x}, \mathbf{l}') = \frac{|\mathbf{x}'^{\top} \mathbf{F} \mathbf{x}|}{\sqrt{l'^2_0 + l'^2_1}} \tag{12}$$

where $\mathbf{l}' = \mathbf{F}^{\top} \mathbf{x}'$ denotes the epipolar line in the first view.

By applying symmetric measurements on the point-epipolar line distances, the energy function defined by Eq. (10) is now revised as follows:

$$\varphi_i(\mathbf{x}_i, \mathbf{x}'_i; \hat{\xi}) = \delta^2(\mathbf{x}'_i, \hat{\mathbf{F}} \mathbf{x}_i) + \delta^2(\mathbf{x}_i, \hat{\mathbf{F}}^{\top} \mathbf{x}_i) \tag{13}$$

This yields geometric errors in pixel locations.

A noise-tolerant variant is to treat the correspondence $\mathbf{x} \leftrightarrow \mathbf{x}'$ as a deviation from the ground truth $\mathring{\mathbf{x}} \leftrightarrow \mathring{\mathbf{x}}'$. When the differences $\|\mathbf{x} - \mathring{\mathbf{x}}\|$ and $\|\mathbf{x}' - \mathring{\mathbf{x}}'\|$ are believed to be small, the sum of squared mutual geometric distances can be approximated by

$$\delta^2(\mathring{\mathbf{x}}, \mathring{\mathbf{l}}') + \delta^2(\mathring{\mathbf{x}}', \mathring{\mathbf{l}}) \approx \frac{(\mathbf{x}'^{\top} \mathbf{F} \mathbf{x})^2}{l_0^2 + l_1^2 + l'^2_0 + l'^2_1} \tag{14}$$

where $\mathring{\mathbf{l}} = \mathbf{F}\mathbf{x}$ and $\mathring{\mathbf{l}}' = \mathbf{F}^{\top} \mathbf{x}'$ are perfect epipolar lines [20]. This first-order approximation to the geometric error is known as the *Sampson distance* [20]. When such a metric is adopted for evaluating egomotion, Eq. (13) is formulated as follows:

$$\varphi_i(\mathbf{x}_i, \mathbf{x}'_i; \hat{\xi}) = \sum_i \frac{(\mathbf{x}'^{\top}_i \hat{\mathbf{F}} \mathbf{x}_i)^2}{(\hat{\mathbf{F}} \mathbf{x}_i)_0^2 + (\hat{\mathbf{F}} \mathbf{x}_i)_1^2 + (\hat{\mathbf{F}}^{\top} \mathbf{x}'_i)_0^2 + (\hat{\mathbf{F}}^{\top} \mathbf{x}'_i)_1^2} \tag{15}$$

As the computation of these epipolar errors only uses 2D correspondences, the energy model can be useful when 3D structures of a scene are not known (i.e. in case of monocular VO). A minimum linear solver, known as the *five-point method*, is credited to Nistér [21]. The absolute scale of $\mathbf{t}$, however, is not possible to be recovered without any reference in the 3D space.

4.2 Perspective Alignment

If the 3D location of a tracked feature is known in the current frame, and its 2D location is found in the next frame, then it is more reliable to use reprojection residuals [17]

$$\varphi_i(\mathbf{g}_i, \mathbf{x}'_i; \hat{\xi}) = \left\| \mathbf{x}'_i - \pi_{\mathbf{K}}(\hat{\mathbf{R}}\mathbf{g}_i + \hat{\mathbf{t}}) \right\|^2_{\boldsymbol{\Sigma}_i} \tag{16}$$

where $\mathbf{g}_i = (X_i, Y_i, Z_i)^\top$ is the current 3D location of the feature, $\mathbf{x}'_i = (x_i, y_i)^\top$ are the feature's image coordinates in the next frame, and $\pi_{\mathbf{K}} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection function that maps a 3D point into the 2D image coordinates using the camera matrix $\mathbf{K}$; $\boldsymbol{\Sigma}_i$ is the 2×2 error covariance matrix of the i-th correspondence. When the backward correspondences $\mathbf{g}'_i \leftrightarrow \mathbf{x}_i$ are known, Eq. (16) can be modelled in inverse mapping form as follows:

$$\varphi_i(\mathbf{g}'_i, \mathbf{x}_i; \hat{\xi}) = \left\| \mathbf{x}_i - \pi_{\mathbf{K}} \left(\hat{\mathbf{R}}^\top \left(\mathbf{g}'_i - \hat{\mathbf{t}} \right) \right) \right\|^2_{\boldsymbol{\Sigma}_i} \tag{17}$$

The geodesic reprojection error has been a widely adopted energy model in camera resectioning, including calibration, pose estimation, and bundle adjustment. Its closed-form linear solution has been extensively studied in the domain of *perspective-from-n-points* (PnP) problems. A popular solver is owed to an efficient algorithm worked out by Lepetit et al. [18]. A linear solution is usually iteratively refined using a derivative-based minimiser (e.g. a Gauss-Newton algorithm). It has been shown that the perspective alignment can be further regularised using the epipolar alignment to reduce the impact of noisy 3D measurements [22].

4.3 Rigid Alignment

If a dense depth map is available and the establishment of 3D point correspondences is straightforward, then a rigid alignment can also be used to measure the fitness of a motion hypothesis.

Given a set of 3D-to-3D correspondences $\mathbf{g_i} \leftrightarrow \mathbf{g}'_i$, where $\mathbf{g_i} = (X_i, Y_i, Z_i)^\top$ and $\mathbf{g'_i} = (X'_i, Y'_i, Z'_i)^\top$. The energy model is defined by

$$\varphi_i(\mathbf{g}_i, \mathbf{g}'_i; \hat{\mathbf{R}}, \hat{\mathbf{t}}) = \left\| g'_i - (\hat{\mathbf{R}} g_i + \hat{\mathbf{t}}) \right\|^2_{\boldsymbol{\Sigma}_i} \tag{18}$$

where $\boldsymbol{\Sigma}_i$ denotes the 3×3 error covariance matrix.

The formulation can be ill-behaved for far points if the 3D coordinates are derived from a disparity map, due to the non-linearity of disparity-to-depth conversion. It is therefore critical to model the covariance matrix properly. If 3D coordinates are obtained by using a two-view triangulation function $\tau : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^3$, then the covariance matrix can be modelled as

$$\boldsymbol{\Sigma} = \mathbf{J}_\tau \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}'} \end{pmatrix} \mathbf{J}_\tau^\top \tag{19}$$

where $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_{x'}}$ are the 2×2 error covariance matrices of image points $\mathbf{x} = (x, y)^\top$ and $\mathbf{x}' = (x', y')^\top$, respectively, and $\mathbf{J}_\tau$ is the 3×4 Jacobian matrix

$$\mathbf{J}_\tau = \left[\tfrac{\partial\tau}{\partial x}(\mathbf{x}, \mathbf{x}') \; \tfrac{\partial\tau}{\partial y}(\mathbf{x}, \mathbf{x}') \; \tfrac{\partial\tau}{\partial x'}(\mathbf{x}, \mathbf{x}') \; \tfrac{\partial\tau}{\partial y'}(\mathbf{x}, \mathbf{x}') \right] \tag{20}$$

with respect to correspondence $\mathbf{x} \leftrightarrow \mathbf{x}'$, used to triangulate $\mathbf{g} = (X, Y, Z)^\top$ [12].

Unlike the previous two alignment models, the rigid model has closed-form solutions that are guaranteed to minimise Eq. (18). A popular choice is based on the quaternion parametrization and *singular value decomposition* (SVD), as shown by Horn [23].

4.4 Photometric Alignment

If image correspondences $\mathbf{x} \leftrightarrow \mathbf{x}'$ are not available at all, one may perform direct photometric matching using a motion hypothesis. In this case, a matching residual is defined over intensity images I and I' as follows:

$$\varphi_i(\mathbf{x}_i, \mathbf{g}_i; \hat{\xi}) = I(\mathbf{x}_i) - I'[\pi_\mathbf{K}(\hat{\mathbf{R}}\mathbf{g}_i + \hat{\mathbf{t}})] \tag{21}$$

assuming that the 3D coordinates $\mathbf{g}_i = (X_i, Y_i, Z_i)$ of image point $\mathbf{x} = (x, y)$ are known. This is also known as *correspondence-free egomotion estimation*, and it is pervasively used by direct VO techniques. Equation (21) can be extended to use a block of pixels instead of single pixel intensities; in this case the subtraction needs be replaced by a proper metric, e.g. the *sum of absolute differences* (SAD).

The evaluation of Eq. (21) is computationally expensive compared to all the aforementioned geodesic criteria. It invokes a rigid transformation, a perspective transformation, and two image sub-sampling procedures. As the minimum of photometric errors can only be approached iteratively, such an expensive cost function will need to be invoked repeatedly for constructing numerically the Jacobian matrix. To ease the incurred burden, many direct methods adopt an inverse compositional form of the residual term [8]

$$\varphi_i{}^k(\mathbf{g}_i, \xi_k; \mathit{\Delta}\xi) = I\left[\pi_\mathbf{K}\Big(\mathit{\Delta}\mathbf{T} \cdot \mathbf{g}_i\Big)\right] - I'\left[\pi_\mathbf{K}\Big(\mathbf{T}_k(\xi_k) \cdot \mathbf{g}_i\Big)\right] \tag{22}$$

where $\mathit{\Delta}\mathbf{T} = \exp(\mathit{\Delta}\xi)$ and $\mathbf{T}_k = \exp(\xi_k)$. Along with the inverse form of

$$\xi_{k+1} = \mathit{\Delta}\xi_k^{-1} \circ \xi_k \tag{23}$$

the Jacobian of $\mathit{\Phi}$ can be written in the chained form

$$\frac{\partial\varphi}{\partial\mathit{\Delta}\xi}(\mathbf{g}_i, \xi_k) = \frac{\partial I}{\partial\mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}_i} \cdot \frac{\partial\pi}{\partial\mathbf{g}}\bigg|_{\mathbf{g}=\mathbf{g}_i} \cdot \frac{\partial\mathbf{T}}{\partial\xi}\bigg|_{\xi=\mathbf{0}} \cdot \mathbf{g}_i \tag{24}$$

which is independent of the current motion hypothesis ξ_k.

The first term of Eq. (24) is the gradient of base image I at key point $\mathbf{x}_i$, which requires only one-time evaluation at the beginning of the minimisation procedure; the second and third term can be calculated symbolically, and the last term is constant, for each tracked key point.

5 Multi-objective Visual Odometry

In this section we describe a feature-based VO approach that uses multiple objectives to achieve egomotion estimation.

5.1 Feature Tracking

Given image I_k of new Frame k, a set of features is detected. Let $\mathcal{F}_k$ be the feature set, and $\nu(\chi)$ be the vector representation of a feature $\chi \in \mathcal{F}_k$. The tracking is performed by finding that feature $\chi' \in \mathcal{F}_{k-1}$ that minimises the distance $\delta\big(\nu(\chi)-\nu(\chi')\big)$, where the similarity metric δ can either be the Hamming distance if ν gives a binary string, or the L_2-norm in any other case.

We also carry out a symmetry check by performing backward feature matching. Any feature χ' is rejected that is not mapped to the original matcher χ. In addition to this check, ambiguity matches are also removed by checking the ratio of distances between (χ, χ') and (χ, χ'') where $\chi'' \in \mathcal{F}_{k-1}$ is the second best match of χ.

5.2 Construction of a Data Term

The tracked features (χ, χ') are then used to construct multiple sets of correspondences. Let $\rho_k(\chi)$ be the image coordinates of a feature observed in Frame k. The 2D-to-2D correspondences $\rho_{k-1}(\chi) \leftrightarrow \rho_k(\chi)$, where $\rho_k(\chi) = \rho_k(\chi')$ is set to the matched feature's location, are used to build the epipolar constraints, denoted by $\mathcal{M}_{\mathrm{EPI}}$. Let $\bar{g}_k(\chi)$ be the recursively filtered 3D coordinates of feature χ in Frame k. The correspondences $\bar{g}_{k-1}(\chi) \leftrightarrow \rho_k(\chi)$ are used to build the projection constraints, denoted by $\mathcal{M}_{\overrightarrow{\mathrm{RPE}}}$.

We also use the intensity-3D-intensity correspondences $I_{k-1}\big[\rho_{k-1}(\chi)\big] \leftrightarrow \bar{g}_{k-1}(\chi) \leftrightarrow I_k\big[\rho_k(\chi)\big]$ to instantiate a set of photometric constraints, denoted by $\mathcal{M}_{\mathrm{PHOTO}}$. Finally, if the measure $g_k(\chi)$ of a feature's 3D coordinates in the new Frame k is available (either from a disparity map, a LiDAR scan, or any other sources), we construct a set of 3D-to-3D constraints $\bar{g}_{k-1}(\chi) \leftrightarrow g_k(\chi)$, and have it denoted by $\mathcal{M}_{\mathrm{RIGID}}$. We also make use of constraints $g_k(\chi) \leftrightarrow \rho_{k-1}(\chi)$ to build backward reprojection constraints $\mathcal{M}_{\overleftarrow{\mathrm{RPE}}}$.

5.3 Outlier Rejection

All the constructed mappings $\mathcal{M}_{\mathrm{EPI}}$, $\mathcal{M}_{\overrightarrow{\mathrm{RPE}}}$, $\mathcal{M}_{\overleftarrow{\mathrm{RPE}}}$, $\mathcal{M}_{\mathrm{RIGID}}$, and $\mathcal{M}_{\mathrm{PHOTO}}$ are collaboratively used in a RANSAC process to remove outliers. First, at the estimation stage a minimum set of correspondences is randomly drawn from one of the five mappings. From the samples, an initial motion hypothesis $\bar{\xi}$ is solved by the closed-form solver associated with the chosen mapping. At the verification stage, the hypothesis is applied to each class.

By means of the appropriate energy model, introduced in Sect. 4, the data terms are evaluated and the inliers are found as the correspondences achieving an error below a pre-defined class-specific threshold. If the population of inliers, summed over all the classes, achieves a record high, the hypothesis is taken as the best model. Such a process goes until a stopping criterion is met.

The closed-form solver and associated energy model for each class are summarised in Table 1. Note that the $\mathcal{M}_{\text{EPI}}$ class is excluded at the estimation stage, as the translation of estimated motion $\bar{\xi}$ does not have an absolute unit, prohibiting it from being used to evaluate data terms in other classes. The $\mathcal{M}_{\text{PHOTO}}$ class is also excluded because there is no closed-form solver for the photometric alignment problem.

Table 1. Summary of data terms (read "Int.-to-int." as "Intensity-to-intensity")

	$\mathcal{M}_{\text{EPI}}$	$\mathcal{M}_{\overrightarrow{\text{RPE}}}$, $\mathcal{M}_{\overleftarrow{\text{RPE}}}$	$\mathcal{M}_{\text{RIGID}}$	$\mathcal{M}_{\text{PHOTO}}$
Alignment	Epipolar	Perspective proj.	Rigid	Photometric
Correspondence	2D-to-2D	3D-to-2D	3D-to-3D	Int.-to-int.
Closed-form solution	Five-point [21]	EPnP [18]	SVD [23]	N/A
Closed-form data needed	5	6	4	N/A
Energy function	Φ_{EPI}	Φ_{RPE}	Φ_{RIGID}	Φ_{PHOTO}

5.4 Unified Energy Models

The best-fit model $\bar{\xi}$ from the RANSAC process serves as the initial guess at the non-linear optimisation stage over the integrated energy model

$$\Phi(\xi) = \Phi_{\text{RPE}}(\xi) + \Phi_{\text{EPI}}(\xi) + \Phi_{\text{RIGID}}(\xi) + \Phi_{\text{PHOTO}}(\xi) \tag{25}$$

with each sub-objective instantiated from the inliers of the corresponding classes. Note that the combination does not use per-class weightings as the residuals are already normalised by the estimated error covariance, as discussed in Sect. 4. Function Φ is minimised using the iterative process described in Sect. 3.2.

6 Experiments

We selected a road scene from the KITTI benchmark suite [24] for evaluating the unified model. The vehicle had travelled 350 m in 188 frames. The VO implementation computes depth maps using OpenCV's SGBM functions and tracks SURF features through the image sequence. The images from the right camera are only used for disparity generation. The depth of a tracked feature is continuously integrated using a recursive Bayesian filter. No additional optimisation technique (e.g. bundle adjustment, lost feature recovery, or similar) has been deployed.

To test how each energy model affects the VO process, we tried all the possible combinations by enabling a subset of data terms among $\mathcal{M}_{\text{EPI}}$, $\mathcal{M}_{\overleftarrow{\text{RPE}}}$, $\mathcal{M}_{\text{RIGID}}$ and $\mathcal{M}_{\text{PHOTO}}$, for each test. Note that the forward-projection term $\mathcal{M}_{\overrightarrow{\text{RPE}}}$ is

Table 2. Egomotion estimation drifts (%) of different energy model combinations. Letters B, P, R, and E, respectively, indicate the use of backward projection ($\mathcal{M}_{\overleftarrow{\mathrm{RPE}}}$), photometric ($\mathcal{M}_{\mathrm{PHOTO}}$), rigid ($\mathcal{M}_{\mathrm{RIGID}}$), and epipolar ($\mathcal{M}_{\mathrm{EPI}}$) alignment models. Maximum and minimum value in each column are in bold.

Model	Best	Worst	Mean	Std.	Model	Best	Worst	Mean	Std.
xxxx	4.97	5.54	5.21	0.27	Bxxx	**5.14**	5.99	5.41	0.34
xPxx	2.26	2.76	2.52	0.21	BPxx	1.99	2.50	2.23	0.21
xxRx	4.65	5.09	4.88	0.15	BxRx	5.10	**6.00**	**5.58**	**0.37**
xPRx	**1.84**	2.39	2.18	0.26	BPRx	1.96	2.56	**2.16**	0.26
xxxE	2.27	2.31	2.28	**0.01**	BxxE	2.21	**2.29**	2.24	0.03
xPxE	2.24	2.71	2.47	0.17	BPxE	2.17	2.48	2.31	0.11
xxRE	2.29	2.38	2.34	0.03	BxRE	2.18	2.31	2.24	0.05
xPRE	2.41	2.59	2.50	0.08	BPRE	2.21	2.40	2.33	0.08

always used as it is required to properly bootstrap the RANSAC process. This results in sixteen configurations. Due to the randomness introduced into the outlier-rejection stage, we carried out five trials for each configuration. Table 2 summarises motion drifts from eighty estimated trajectories.

In most cases, using additional energy model(s) significantly reduces the drift in estimated egomotion. Exceptions are observed in the cases of xxRx, Bxxx, and BxRx.

When the rigid alignment is solely imposed (xxRx), the drift only slightly reduces by 0.35%. A result, worse than the forward-projection-only baseline configuration (xxxx), is found in BxRx where it is simultaneously applied with the back-projection alignment model. A similar result is observed when the backward projection is used (Bxxx). The loss of accuracy can be due to the use of feature depths obtained in frame $t+1$, where the recursive Bayesian filter has not been applied, as the egomotion from t to $t+1$ is not yet estimated (see Fig. 1).

Interestingly the results show that, when the epipolar term is used with the forward projection model (xxxE), the VO process yields highly robust estimates with a very low standard deviation (0.01%). Such finding corresponds to the authors' previous work [22].

We further compared the baseline model with the best case, worst case, and the case using all energy models. Accumulated drifts are plotted in Fig. 2. In the best case the accuracy is improved by 58.3%, while by imposing all four energy models, this option achieves an improvement by 54.8%.

We also profiled the run-time of each test case. The processing time for each frame are 230 ms and 200 ms for the best case and the baseline implementation, respectively. This time measurement indicates that the introduction of additional energy terms only incurs 13% more computational cost, while an improvement of above 50% in terms of accuracy is attainable.

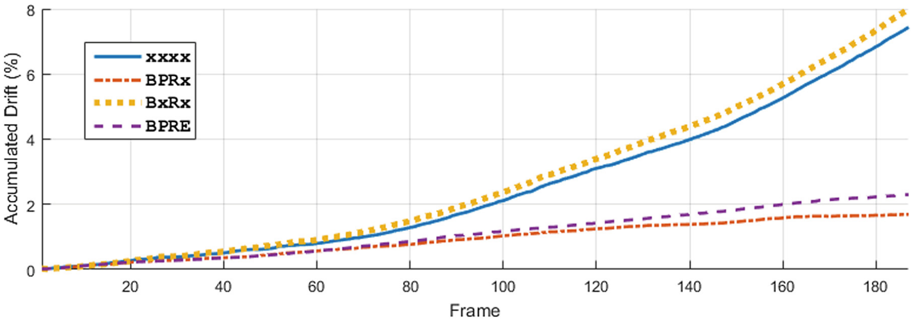

Fig. 2. Drift analysis of the best (BPRx), worst (BxRx), all-enabled (BPRE), and the baseline model (xxxx)

7 Conclusions

We reviewed four energy models, pervasively used in the context of VO, and formulated a unified model. Based on the model, our implementation deploys a multi-class RANSAC strategy to remove outliers with a proven enhanced robustness. Real-world experimental results show that, by taking into account multiple objectives, egomotion estimation is significantly improved over the traditional forward-projection model, at an affordable minor increase in computational costs.

In future work we test the proposed multi-modal VO technique in a wider range of road scenes. It is also interesting to study the interference between alignment models for the phenomena that the all-enabled combination does not achieve the best performance in the tested sequence.

References

1. Scaramuzza, D., Fraundorfer, F.: Visual odometry: part I - the first 30 years and fundamentals. IEEE Robot. Autom. Mag. **18**, 80–92 (2011)
2. Maimone, M., Cheng, Y., Matthies, L.: Two years of visual odometry on the Mars exploration rovers. J. Field Robot. **24**(3), 169–186 (2007)
3. Irani, M., Anandan, P.: All about direct methods. In: Proceedings of ICCV Workshop Vision Algorithms: Theory Practice, pp. 267–277 (1999)
4. Baker, S., Mathews, I.: Equivalence and efficiency of image alignment algorithms. In: Proceedings of International Conference on Computer Vision Pattern Recognition, vol. 1, pp. 1090–1097 (2001)
5. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proceedings of International Symposium on Mixed Augmented Reality, pp. 1–10 (2007)
6. Davison, A., Reid, I., Molton, N., Stasse, O.: MonoSLAM: real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 1052–1067 (2007)
7. Newcombe, R., Lovegrove, S., Davison, A.: DTAM: dense tracking and mapping in real-time. In: Proceedings of IEEE International Conference on Computer Vision, pp. 2320–2327 (2011)

8. Forster, C., Pizzoli, M., Scaramuzza, D.: SVO: fast semi-direct monocular visual odometry. In: Proceedings of IEEE International Conference on Robotics Automation, pp. 15–22 (2014)
9. Engel, J., Sturm, J., Cremers, D.: Semi-dense visual odometry for a monocular camera. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1449–1456 (2013)
10. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: large-scale direct monocular SLAM. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 834–849. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_54
11. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Trans. Pattern Anal. Mach. Intell. (99) (2017)
12. Maimone, M., Cheng, Y., Matthies, L.: Two years of visual odometry on the Mars exploration rovers. J. Field Robot. Special Issue Space Robot. Part I **24**, 169–186 (2007)
13. Kitt, B., Geiger, A., Lategahn, H.: Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In: Proceedings of IEEE Intelligent Vehicles Symposium, pp. 486–492 (2010)
14. Badino, H., Yamamoto, A., Kanade, T.: Visual odometry by multi-frame feature integration. In: Proceedings of International ICCV Workshop Computer Vision Autonomous Driving (2013)
15. Chien, H.-J., Geng, H., Chen, C.-Y., Klette, R.: Multi-frame feature integration for multi-camera visual odometry. In: Bräunl, T., McCane, B., Rivera, M., Yu, X. (eds.) PSIVT 2015. LNCS, vol. 9431, pp. 27–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29451-3_3
16. Mur-Artal, R., Tardos, J.: ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. arXiv preprint arXiv:1610.06475 (2016)
17. Nister, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proceedings of International Conference on Computer Vision Pattern Recognition, pp. 652–659 (2004)
18. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: an accurate O(n) solution to the PnP problem. Int. J. Comput. Vis. **81**, 155–166 (2009)
19. Levenberg, K.A.: Method for the solution of certain non-linear problems in least squares. Q. Appl. Math. **2**, 164–168 (1944)
20. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)
21. Nister, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 756–777 (2004)
22. Chien, H.-J., Klette, R.: Regularised energy model for robust monocular egomotion estimation. In: Proceedings of International Joint Conference on Computer Vision Imaging Computer Graphics: Theory Applications, vol. 6, pp. 361–368 (2017)
23. Horn, B.: Closed-form solution of absolute orientation using unit quaternions. J. Opt. Soc. Am. A **4**, 629–642 (1987)
24. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)