Exploiting Motion Priors in Visual Odometry for Vehicle-Mounted Cameras with Non-holonomic Constraints

Davide Scaramuzza, Andrea Censi, Kostas Daniilidis

Abstract— **This paper presents a new method to estimate the relative motion of a vehicle from images of a single camera. The biggest problem in visual motion estimation is data association; matched points contain many outliers that must be detected and removed so that the motion can be estimated accurately. A very established method for robust motion estimation in the presence of outliers is the five-point RANSAC algorithm. Five-point RANSAC operates by generating motion hypotheses from randomly-sampled minimal sets of five-point correspondences. These hypotheses are then tested against all data points and the motion hypothesis that after a given number of iterations returns the largest number of inliers is taken as the solution to the problem. A typical drawback of RANSAC is that the number of iterations required to find a suitable solution grows exponentially with the number of outliers, often requiring thousands of iterations for typical data from urban environments. Another problem is that – due to its random nature – sometimes the found solution is not the "best" solution to the motion estimation problem. In this paper, we describe an algorithm for relative motion estimation in the presence of outliers, which does not rely on RANSAC. Contrary to RANSAC, motion hypotheses are not generated from randomly-sampled point correspondences, but from a "proposal distribution" that is built by exploiting the vehicle non-holonomic constraints. We show that not only is the proposed algorithm significantly faster than RANSAC, but that the returned solution may also be better in that it favors the underlying motion model of the vehicle, thus overcoming the typical limitations of RANSAC. Additionally, the proposed algorithm provides the likelihood of the motion estimate, which can be very useful in all those applications where a probability distribution of the position of the vehicle is required (e.g., SLAM). Finally, the performance of the proposed method is compared to that of the standard five-point RANSAC on real images collected from a vehicle moving in a cluttered, urban environment.**

I. Introduction

Visual odometry is the problem of estimating the ego-motion of a vehicle from onboard-camera images. Several works have been recently produced using both stereo or monocular cameras [1]–[6]. Basically, visual odometry operates by incrementally computing the motion between consecutive frames. This is done by extracting salient points (such Harris, FAST, SIFT, etc.) from both images and matching them according to some similarity measure. However, matched points are usually contaminated by outliers, that is, wrong data associations. Outliers must be carefully removed so that the motion can be estimated accurately.

The *random sample consensus* (RANSAC) [7] has been established as the standard method for motion estimation in the presence of outliers. In RANSAC, the motion ($\mathbf{R}$,$\mathbf{T}$) is estimated from feature correspondences. The idea behind RANSAC is to compute model hypotheses from randomly-sampled minimal sets of data points and then verify these hypotheses on the other data points. The hypothesis that shows the highest consensus with the other data is selected as solution. The number of iterations N that is necessary to guarantee that a correct solution is found can be computed as

$$N = \frac{\log(1-p)}{\log(1-(1-\varepsilon)^s)}, \tag{1}$$

where ε is the percentage of outliers in the data points, p is the requested probability of success and s is the number of data points necessary for estimating the model. For unconstrained motion (6DoF) of a calibrated camera this would be 5 correspondences [8]. This made the 5-point RANSAC (1) the standard algorithm for unconstrained motion estimation in the presence of outliers.

The drawback in RANSAC is that the number of iterations grows exponentially in the number of outliers (see (1)). In some cases, the 5-point RANSAC can require up to one thousand iterations for typical data from a vehicle in urban environments. Because of this, several works have been produced in the endeavor of reducing the number of iterations. In [9], the authors manage to do it by ranking the correspondences based on their similarly. In [10], the authors use a preemptive scoring of the motion hypotheses. Finally, in [11] the authors incorporate feature uncertainty and show that this determines a decrease in the number of potential outliers, enforcing thus a reduction in the number of iterations. What all these methods have in common is that the motion hypotheses are still generated from point correspondences, which is an expensive test as it may involve SVD and Groebner-basis decompositions. In addition, for each candidate set of data points 5-point RANSAC returns up to ten motion solution, each to be tested. An alternative algorithm was proposed in [12] for EKF-based visual odometry. They proposed to use the available prior probabilistic information from the EKF in the RANSAC model-hypothesize stage.

In this paper, we propose a novel algorithm to compute both the relative motion of a single camera and wrong data associations, which does not rely on a RANSAC-scheme. We pose the problem as a maximum-likelihood estimation. The algorithm operates by estimating a proposal distribution that *captures* the main components of the motion, and

Davide Scaramuzza is with the GRASP Lab, department of Computer and Information Science, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, USA. Kostas Daniilidis is director of the GRASP Lab, University of Pennsylvania. Andrea Censi is with the Control and Dynamical Systems department, Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, USA

which is based on the vehicle non-holonomic constraints. We show that sampling from this proposal is equivalent to computing the joint posterior probability of the complete motion. Alternatively, if one is looking for the maximum-likelihood solution, then one can choose the solution that has given more inliers. We show that the proposed algorithm is significantly faster than the 5-point RANSAC and is also more accurate in that it favors the underlying motion model of the vehicle, thus overcoming the typical limitations of RANSAC (i.e., motion error due to the different "quality" of the inliers).

Note that this paper is an extension of our previous work on the 1-point RANSAC algorithm [13], [14]. However, the difference with that work is that here we are relaxing the constraint of planar and circular motion.

The paper is structured as follows. In section II we provide a Bayesian perspective of the motion estimation problem. In section III, we describe how to compute the proposal distribution using the vehicle non-holonomic constraints. In section IV, we explain how to compute the motion prior. In section V, we detail our algorithm. Finally, in sections VI and VII, we present the experimental results and draw the conclusions.

II. Probabilistic definition of relative-motion estimation

In this section we describe the relative-motion estimation problem from a Bayesian perspective.

We assume that a feature detection and matching procedure has already been done. Therefore, we have a set $X = \{\mathbf{x_0}, \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ of n image points ($x_i \in \mathbb{R}^2$), seen by the camera at the first position, and the set $X' = \{\mathbf{x'_0}, \mathbf{x'_1}, \mathbf{x'_2}, ..., \mathbf{x'_n}\}$ of the corresponding image points, seen by the same camera at the next position. We assume that these correspondences are given but are not certain; therefore we say that $\mathbf{x_i}$ is the *putative* image correspondence of $\mathbf{x'_i}$. We assume that any information about the feature appearance has already been used, and only the motion of the vehicle can disambiguate between *inliers* and *outliers*. We define a set of binary hidden variables $\{\alpha_i\}_{i=1}^n$ such that $\alpha_i = 1$ if the i-th correspondence is an inlier, and 0 otherwise.

Let $\mathbf{R} \in SO(3)$ and $\mathbf{T} \in \mathbb{R}^3$ represent the unknown relative motion between the two camera positions. It is well known that with a single camera we can only recover the direction of the translation and not its length. Therefore, for convenience we impose $\|\mathbf{T}\| = 1$. The motion is therefore described by five parameters.

X and X' represent the measured *data*, while $\mathbf{R}$ and $\mathbf{T}$ are the quantities that we want to *estimate* from the data. Writing this in probabilistic terms, the relative-motion estimation problem consists in estimating the joint posterior probability over $\mathbf{R}$ and $\mathbf{T}$ from the data X and X', that is

$$p(\mathbf{R}, \mathbf{T} | X, X'). \tag{2}$$

When posed as a maximum-likelihood estimation problem, the solution to the motion-estimation then becomes that of finding the combination of $\mathbf{R}$ and $\mathbf{T}$ that maximize (2). An alternative solution consists in drawing samples from (2) if it must be used as a proposal distribution in a SLAM algorithm.

As a measure of the *likelihood* of a given motion, the number of inliers that support the motion is generally adopted [15]. In these terms, maximizing (2) becomes equivalent to finding the motion that maximizes the number of inliers.[1]

A brute-force approach to solve (2) would be to perform a full search over the five-dimensional space of motion parameters. This is, however, computationally unfeasible.

A more practical solution to (2) is via random sample consensus (i.e., RANSAC [7]). In RANSAC, motion hypotheses are generated by drawing randomly minimal sets of point correspondences (e.g., five correspondences, see [8]). These hypotheses are then tested against all the point correspondences and the motion hypothesis that – after a given number of iterations – returns the largest number of inliers is taken as the solution to the problem.[2] A typical problem with RANSAC is that the minimum number of iterations required to return a set of points free out outliers within a given confidence grows exponentially with the percentage of outliers in the data. Another problem of RANSAC is that, sometimes, the best found solution – i.e., the one with the largest number of inliers - is not the best solution to the motion estimation problem [16]. Indeed, inliers do not all have the same "quality". Some inliers are better to estimate rotation than translation, or vice versa. As shown in [4], far-distance points are good for estimating rotation, while close-distance points are optimal for estimating translation. This means that, depending on the proportion of far and close points in the data, the inliers found by RANSAC might be different, and so the final motion estimate. Additionally, as studied in [16], the solution of RANSAC is influenced by the resolution of the camera, becoming more evident for omnidirectional cameras.

Contrary to RANSAC, in this paper we propose to select the motion hypotheses not from randomly-drawn minimal sets of image correspondences but directly from a proposal distribution of the motion which is obtained by exploiting the non-holonomic constraints of the vehicle.

Let us parametrize the rotation $\mathbf{R}$ in terms of its *yaw* (θ), *pitch* (β), and *roll* (γ) angles and $\mathbf{T}$ in terms of its azimuth (ϕ) and elevation (δ) angles. Using this angular parametrization, (2) can be rewritten as

$$p(\theta, \beta, \gamma, \phi, \delta | X, X'). \tag{3}$$

We call this distribution the *target distribution* as it is the one that we want to compute.

[1] We recall that the inliers represent a subset of the all image correspondences, for which the reprojection error is smaller than a user-specified threshold (typically 1 pixel). The reprojection error is computed by, first, triangulating the image points in the 3D space using the knowledge of the motion, and, then, reprojecting the 3D feature into the images. The reprojection error is defined as the distance in pixels between the measured image point and the reprojected 3D point. More details can be found in [15].

[2] The motion is then refined by including all the inliers generated by this motion hypothesis.

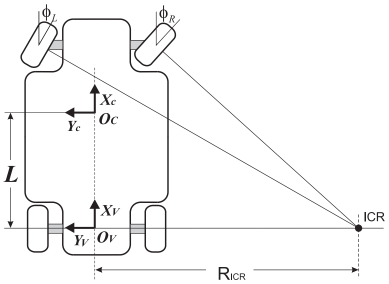


Fig. 1. General Ackermann steering principle.

If we now apply the definition of conditional probability, we can expand (3) as:

$$p(\theta,\beta,\gamma,\phi,\delta|X,X') = p(\theta|X,X')p(\beta,\gamma,\phi,\delta|\theta,X,X') \quad (4)$$

We will refer to $p(\theta|X,X')$ as the *proposal distribution* and to $p(\beta,\gamma,\phi,\delta|\theta,X,X')$ as the *motion prior*.

In the next section, we will describe how to compute the proposal distribution and the motion prior from image correspondences. In particular, we will show that the proposal distribution can be obtained directly and very efficiently from image correspondences by exploiting the non-holonomic constraints of the vehicle.

III. Computing the proposal distribution for θ

For a wheeled vehicle to exhibit rolling motion, a point must exist around which each wheel of the vehicle follows a circular course [17]. This point is known as Instantaneous Center of Rotation (ICR) and can be computed by intersecting all the roll axes of the wheels (Fig. 1). This property holds for any robot, and in particular for car-like and differential-drive. For cars the existence of the ICR is ensured by the Ackermann steering principle [17]. This principle ensures a smooth movement of the vehicle by applying different steering angles to the inner and outer front wheel while turning (see Fig. 1).

As the reader can perceive, the motion of a camera fixed on the vehicle can then be locally described with circular motion.[3] Notice that this constraint also reduces the degrees of freedom of the motion to two, namely the rotation angle θ and the radius of curvature. As will see, this constraint allows us to compute directly the proposal distribution $p(\theta|X,X')$ from the point correspondences.

A. Incorporating vehicle non-holonomic constraints into the camera motion

Let us assume that the camera is fixed somewhere on the vehicle (with the origin in O_C, Fig. 2) with the axis $\mathbf{z_C}$ orthogonal to the plane of motion and $\mathbf{x_C}$ oriented perpendicularly to the back wheel axis.[4]

The origin O_V of the vehicle reference frame can be chosen arbitrarily. For convenience, we set O_V at the intersection of $\mathbf{x_C}$ with the rear-wheel axis,[5] and $\mathbf{x_V}$ aligned with $\mathbf{x_C}$ (Fig. 2).

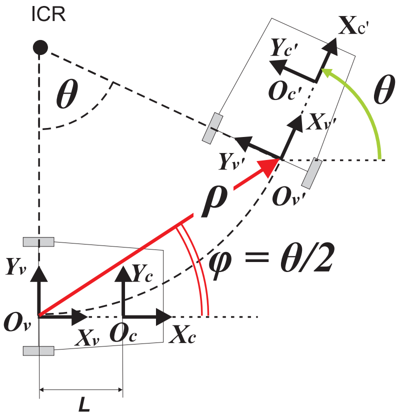


Fig. 2. Relation between camera axes in circular motion.

Following these considerations, the transformation $A_V^C = (\mathbf{R_V^C}, \mathbf{T_V^C})$ from the camera to the vehicle reference system can be written as $\mathbf{R_V^C} = \mathbf{I_{3\times3}}$ and $\mathbf{T_V^C} = [-L,0,0]^T$, where L is the distance between the camera and the back wheel axis (Fig. 2).

If the vehicle undergoes perfect circular motion with rotation angle θ, then the direction of translation ϕ of the vehicle must satisfy the *circular motion constraint*

$$\phi = \theta/2, \quad (5)$$

which can be easily verified by trigonometry. Accordingly, the transformation between the first and the second vehicle position $A_{V'}^V = (\mathbf{R_{V'}^V}, \mathbf{T_{V'}^V})$ can be written as:

$$\mathbf{R_{V'}^V} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \mathbf{T_{V'}^V} = \rho \cdot \begin{bmatrix} \cos(\frac{\theta}{2}) \\ \sin(\frac{\theta}{2}) \\ 0 \end{bmatrix} \quad (6)$$

where ρ is the vehicle displacement (Fig. 2).

Following these considerations, the overall transformation $A_{C'}^C = (\mathbf{R_{C'}^C}, \mathbf{T_{C'}^C})$ between the first and second camera position can be computed as a composition of the following three transformations, that is:

$$A_{C'}^C = A_V^C \circ A_{V'}^V \circ A_{C'}^{V'} = A_V^C \circ A_{V'}^V \circ A_V^{C^{-1}} \quad (7)$$

where we used $A_{C'}^{V'} = A_V^{C^{-1}}$. And from this, we obtain:

$$\mathbf{R_{C'}^C} = \mathbf{R_{V'}^V}, and \quad \mathbf{T_{C'}^C} = \begin{bmatrix} L\cos(\theta) - \rho\cos(\frac{\theta}{2}) - L \\ \rho\sin(\frac{\theta}{2}) - L\sin(\theta) \\ 0 \end{bmatrix}. \quad (8)$$

B. Applying epipolar geometry

We would like to recall some fundamentals of computer vision. Let $\mathbf{x} = [u,v,w]^{\mathrm{T}}$ and $\mathbf{x}' = [u',v',w']^{\mathrm{T}}$ be the normalized image coordinates of a scene point seen from the two camera positions.[6]

[3] Note, rectilinear motion can be represented along a circle with infinite radius of curvature.

[4] Observe that once the camera is installed on the vehicle the axes can be rearranged in the way above with a simple transformation of coordinates.

[5] We observed that by this choice the equations are notably simplified.

[6] With the term *normalized image coordinates* we denote the 3D vectors obtained by back projecting the image points onto a unit sphere with origin on the camera center. This operation is always possible if the camera is calibrated.

As known in computer vision [15], the two unknown camera positions and the image coordinates must verify the epipolar constraint

$$\mathbf{x'}^{\mathrm{T}}\mathbf{E}\mathbf{x} = 0. \tag{9}$$

$\mathbf{E}$ is called *essential matrix* and is defined as $\mathbf{E} = [\mathbf{T}]_{\times}\mathbf{R}$. [7]

This said, we can then compute the essential matrix for our case as $\mathbf{E} = [\mathbf{T}^{\mathbf{C}}_{\mathbf{C'}}]_{\times}\mathbf{R}^{\mathbf{C}}_{\mathbf{C'}}$, that is,

$$\mathbf{E} \;=\; \begin{bmatrix} 0 & 0 & \sin(\frac{\theta}{2}) - \frac{L}{\rho}\sin(\theta) \\ 0 & 0 & \cos(\frac{\theta}{2}) + \frac{L}{\rho}(1-\cos(\theta)) \\ \frac{L}{\rho}\sin(\theta) + \sin(\frac{\theta}{2}) & \frac{L}{\rho}(1-\cos(\theta)) - \cos(\frac{\theta}{2}) & 0 \end{bmatrix}. \tag{10}$$

At this point, notice that if we assume $L = 0$ [8] (10) gets notably simplified and can be rewritten as

$$\mathbf{E} = \begin{bmatrix} 0 & 0 & \sin(\frac{\theta}{2}) \\ 0 & 0 & \cos(\frac{\theta}{2}) \\ \sin(\frac{\theta}{2}) & -\cos(\frac{\theta}{2}) & 0 \end{bmatrix}. \tag{11}$$

Finally, by substituting (11) into the epipolar constraint (9), we obtain the following homogeneous equation that needs to be satisfied by every pair of point correspondences $\mathbf{x}$, $\mathbf{x}'$:

$$\sin\left(\frac{\theta}{2}\right)\cdot(u'w \,+\, w'u) \;+\; \cos\left(\frac{\theta}{2}\right)\cdot(v'w \,-\, w'v) = 0 \tag{12}$$

C. Extracting the proposal distribution

We can see that (12) depends only on the single parameter θ, which can be computed from a single feature correspondence as

$$\theta \;=\; -2\tan^{-1}\left(\frac{v'w \,-\, w'v}{u'w \,+\, w'u}\right). \tag{13}$$

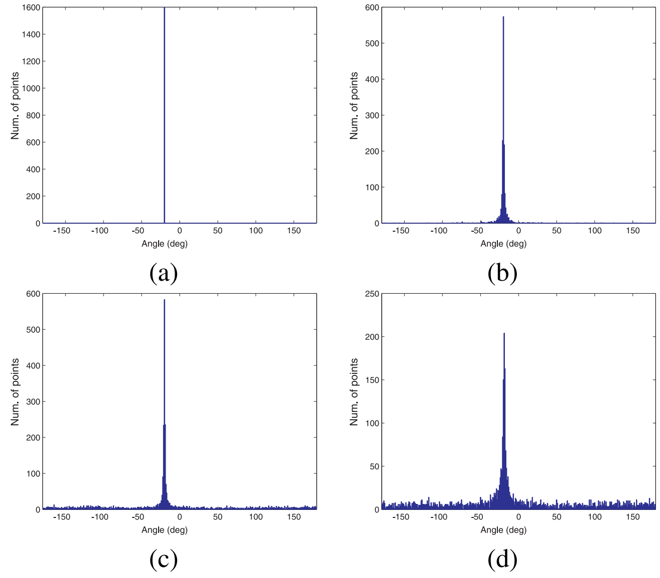
We will refer to (13) as the 1-*point algorithm* [13], [14].

For n feature correspondences we can then build a histogram, where each bin contains the number of points that vote for the same θ. This histogram represents exactly the *proposal distribution* $p(\theta|X,X')$ that we were looking for. As we can see, it is computed directly from the correspondence sets X and X' without requiring any prior motion estimation step or RANSAC scheme.

To recap, equation (13) has been determined from the following assumptions:

- The vehicle motion is planar: i.e., $\beta = 0$, $\gamma = 0$, and $\delta = 0$.
- The vehicle motion is circular: $\phi = \theta/2$ (see equation (5))
- $L/\rho = 0$.

From this we can expect that the shape of the proposal distribution will in general change depending on deviations from the planar-and-circular-motion assumption and from the condition $L/\rho = 0$. Additionally, it will depend on the image noise and on the percentage of outliers in the data.



(a)

(b)

(c)

(d)

Fig. 3. Sample histograms generated from simulated data assuming:
(a) No image noise, no outliers, and $L = 0$.
(b) 0.5 pixel Gaussian noise, no outliers, and $L = 0$.
(c) 0.5 pixel Gaussian noise, 50% outliers, and $L = 0$.
(d) 0.5 pixel Gaussian noise, 50% outliers, $L/\rho = 2$, and non-planar and non-circular motion (we added 5 degrees in roll and pitch for the rotation, 5 degrees in elevation for the translation, and a 10-degree deviation from the circular motion constraint (5)).

We analyzed in simulation the influence of these factors on the proposal distribution. Some sample histograms are depicted in Fig. 3. As we can expect, if the camera motion is both planar and circular and, additionally, there is no image noise and no outliers, all point correspondences will vote for the same θ and the histogram will feature a single bin (Fig. 3a). Conversely, if we just add a 0.5 pixel standard-deviation Gaussian noise to the point correspondences, the distribution broadens (Fig. 3b). Adding 50% outliers in addition to the image noise only increases the tails of the distribution (Fig. 3c). Finally, small deviations from the planar and circular motion assumption and from the condition $L/\rho = 0$ increase the variance of the distribution (Fig. 3d).

This behavior reflects that of the probability distribution we were looking for: the closer we are to the ideal condition of perfectly-planar-circular motion, the narrower the distribution. The farther we are we from the ideal condition, the wider the distribution, and – in other words – the less certain we are about the real motion of the car.

The next question that now we would like to answer is: How well does this proposal distribution represent the true distribution of the true *yaw* angle of the car? To answer this question we need to compare the best estimate of θ obtained from the distribution with ground truth values.

In any of the situations analyzed in Fig. 3 we found that the *median*[9] of the distribution was always very close (within only 0.5 degrees) to the true value of θ. This conjecture was then tested on real data. We computed the proposal

[7] $[\mathbf{T}]_{\times}$ is the skew symmetric matrix $\begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix}$

[8] This is always satisfied when the camera is positioned above the rear-wheel axis or, alternatively, when the ration L/ρ is sufficiently smaller than 1

[9] We would like to remind that for non-Gaussian distributions the *median* (and not the *arithmetic mean*) is the best estimate of the true value.

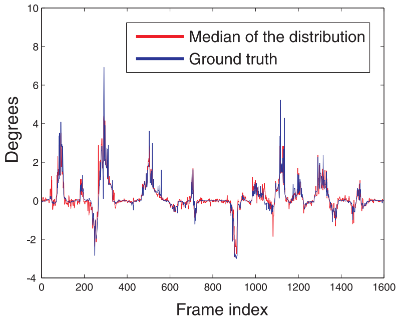


Fig. 4. Comparison between the median of the proposal distribution for θ and ground truth. The ground truth was recorded from GPS, IMU, and wheel odometry.

distribution for a dataset of 15000 images collected from a car in a urban environment. We found that in 99% of the cases the median of the distribution differed from the ground-truth by less than 0.5 degrees. A comparison between the median of the distribution and the ground truth for 1600 frames is depicted in Fig. 4. The ground truth was recorded from GPS, IMU, and wheel odometry.

D. Computational cost

Finally, observe that the proposal distribution just requires to compute (13) n times, where n is the number of putative correspondences. In our experiments we limited n to 4000. The computation of the proposal distribution took less than 1 microsecond for every image pair on a 2GHz Dual Core laptop.

IV. Computing the motion prior

The next step towards the estimation of the target distribution (4) is the computation of the *motion prior*

$$p(\beta,\gamma,\phi,\delta|\theta,X,X'). \tag{14}$$

Recall that θ represents the yaw angle of the rotation, while ϕ the azimuth of the translation vector, and, therefore, they encode the *planar component* of the vehicle motion. Conversely, β and γ represent the roll and pitch angles of the rotation, while δ is the elevation angle of the translation vector. Thus, they encode the *non planar component* of the motion.

As explained in the previous sections, *if the motion was exactly planar* ($\beta,\gamma,\delta = 0$), the parameter θ could be estimated very efficiently using the 1-point algorithm (13), and the kinematic constraints would impose $\phi = \frac{\theta}{2}$.

In practice, the motion of a vehicle in an urban environment is not exactly planar. Therefore, the θ and ϕ estimated assuming planar motion would not be correct. However, the vehicle motion is not very far from planarity either. Therefore, the value of θ estimated using the planarity assumption does have *some* information about the real θ and the other parameters. More precisely, we need an expression for the distribution (14), which we interpret as the distribution of the complete motion given the yaw value (θ) of its planar component.

We can have a direct estimate of the distribution (14) by looking at either the distribution of the solution, or by looking at the ground-truth data obtained by using other sensors, such as IMU. In our experiments conducted in real and diversified urban environments [6], [13], [14], we found two characteristics of that distributions: the planar component is by far the dominant component, meaning that the non-planar components are very small; and the deviations from the planar motion are essentially uncorrelated with the planar components, because they are due mainly to steps, humps, and road irregularities.

These characteristics make it reasonable to assume a Gaussian distribution for the parameters β,γ,ϕ,δ:

$$p(\beta,\gamma,\phi,\delta|\theta,X,X') = p(\beta,\gamma,\phi,\delta|\theta) = \mathcal{N}(\mu,\Sigma),\text{[10]} \tag{15}$$

where μ is the mean of the distribution and Σ its covariance matrix. The mean μ is given by the planar component of the motion:

$$\mu = [\hat{\beta}\ ,\ \hat{\gamma}\ ,\ \hat{\phi}\ ,\ \hat{\delta}] = [0\ ,\ 0\ ,\ \frac{\theta}{2}\ ,\ 0]. \tag{16}$$

For the covariance matrix Σ, we assume that β, γ, ϕ, and δ are uncorrelated[11].

At this point, we can choose the individual variances according to the maximum deviations from the planar and circular motion constraint that we can tolerate. In our experiments, we found that it was good to assume fixed values of

$$\sigma_\beta = \sigma_\gamma = \sigma_\delta = 3\text{deg}.$$

If we accept a 99%-confidence, this means we can tolerate deviations from the planar assumption up to 3σ, that is, up to 9 degrees.

The uncertainty of ϕ, instead, depends on the value of θ. We found in our experiments that the translation direction ϕ never exceeded the ideal value $\theta/2$ by more than $\theta/2$. Again, assuming that $3\sigma_\phi = \theta/2$, we chose

$$\sigma_\phi = \theta/6.$$

Finally, we note that some of the considerations for urban environments would fail should the vehicle be driving in extreme conditions (inclined racetracks, off-road driving). However, notice that the algorithm degrades gracefully, as we could take care of those conditions by increasing the variance parameters. More samples would be needed in the next step, but, as long as the planar component of the motion has some relevance to the actual motion, there still would be an advantage in using this proposal distribution for the search.

[10] Notice that we removed X and X' because the motion prior does not depend on the data points.

[11] This assumption does not cause a decrease of the performance. By contrary, it means that we are less certain about the motion and that, consequently, we have more choice in selecting the good motion parameters. Indeed, when variables are uncorrelated the Gaussian distribution will appear less “stretched” than when variable are correlated.

V. The complete algorithm

A. Estimating the motion solution and the joint posterior

At this point, we have all the pieces ready to describe the complete method, which is shown as Algorithm 1. We name our algorithm MOBRAS (MOdel Based RAndom Sampling).

The basic idea is to use the efficient 1-point algorithm (described in Section III) and the knowledge of the motion prior (described in Section IV) to be able to generate guesses of the complete motion by sampling. More in detail: one samples a random correspondence pair. From the two points, one computes θ^k under the assumption of planar motion using (13).[12] Then, using the motion prior, one can obtain a sample of the other parameters $\beta^{(k)}, \gamma^{(k)}, \phi^{(k)}, \delta^{(k)}$.

Given a guess for the motion, we can easily distinguish inliers from outliers using the reprojection error (this was summarized in footnote 1); this means, we can compute deterministically the variables $\{\alpha_i^{(k)}\}_{i=1}^n$. Now, given the inliers, we can compute the refined complete motion $(\mathbf{R}^{(k)}, \mathbf{t}^{(k)})$ very efficiently and accurately using least squares (for this, we used the algorithm in [8]).

The distribution $(\mathbf{R}^{(k)}, \mathbf{t}^{(k)})$ represents the answer to the visual odometry problem (2). If only one answer is needed – that is, we are looking for the maximum likelihood solution – then one can choose the solution that has given more inliers.

Algorithm 1 MOBRAS

```
Repeat $N$ times:
  1) Sample from $p(\theta|X, X')$:
      a) Sample a random feature correspondence $k$.
      b) Compute $\theta^{(k)}$ under the assumption of planar motion using (13).
  2) Sample $\beta^{(k)}, \gamma^{(k)}, \phi^{(k)}, \delta^{(k)}$ from the motion prior $p(\beta, \gamma, \phi, \delta|\theta)$.
  3) Given a guess for the complete motion, compute inliers and outliers: $\{\alpha_i^{(k)}\}_{i=1}^n$ = compute-inliers$(\beta^{(k)}, \gamma^{(k)}, \phi^{(k)}, \delta^{(k)}, X, X')$
  4) Recompute the optimal motion $(\mathbf{R}^{(k)}, \mathbf{t}^{(k)})$ using least-squares on the inliers.
```

B. Remarks

We are sampling from a 5-dimensional distribution, thus one might think intuitively that this would be a very inefficient method as many samples would be needed to cover a 5-dimensional Gaussian distribution. However, there are several considerations to be made. Firstly, remember that the assumption is that the motion is almost planar, and that we compute exactly the planar component of the motion, which serves as the mean of the distribution. Therefore, we use sampling only to capture the unmodelled effects, while effectively we solve analytically for the main component of the motion. Thus the distribution is very compact.

Moreover, notice that we do not need to cover the whole distribution to capture all possible hypotheses: the samples are still only guesses for the whole motion that are used to choose the inliers. In fact, the final motion is computed from the inliers, forgetting the motion guess. Thus, we only need to sample as many guesses as necessary to be able to capture the ambiguities in the choice of the inliers. In the next section, we will show that using 100 samples is more than sufficient to compute accurately the relative motion for typical data in urban environments.

Finally, one might erroneously think that N iterations of our algorithm cost as much as N iterations of RANSAC. This is not true. In fact, in RANSAC motion hypotheses are generated from minimal sets of *data points*, while in our algorithm motion hypotheses are generated directly from the *proposal distribution*, thus avoiding the expensive step of computing them from the data points. In fact, notice that for each candidate point set the 5-point RANSAC returns up to ten motion solutions and this involves both SVD and Groebner-basis decompositions. Another reason our algorithm is faster than RANSAC is that the motion hypotheses are not generated at random but from both a proposal distribution and a motion prior which take into account the dominant planar component of the motion. In the standard 5-point RANSAC this prior information is not used and therefore the number of iterations grows (exponentially) with the number of outlier. Conversely, we can always use the same number of samples regardless of the number of outliers in the data.

VI. Results

The method described in this paper was successfully tested on an 15000-image dataset collected from a real vehicle moving in a urban environment (Fig. 5).

To show the generality of application of our approach, the images were taken with an omnidirectional camera. The camera used was a SONY XCD-SX910 – image size 640 × 480 pixels – equipped with a panoramic hyperbolic mirror from EyeSee360.

For ground-truth data acquisition, the vehicle was equipped with a GPS, an inertial measurement unit (IMU), and wheel encoders. The position and orientation of the vehicle was computed by means of an extended Kalman filter as described in [18].

For feature extraction, we used the Harris detector [19].

The performance of the proposed algorithm is compared to that of the 5-point RANSAC, which is considered the standard in visual odometry [1]. The performance is evaluated in terms of accuracy of the estimated motion, number of samples, and execution time.

A. Motion accuracy

There are two different criteria to evaluate the motion accuracy. The first one is by comparing the estimated trajectory with ground-truth data. The second one is by comparing the estimate of the relative motion between consecutive frames with ground-truth data. With the first criterion, however, one

[12] Notice that this is perfectly equivalent to sampling from the proposal distribution. The advantage of doing so is that we avoid explicitly computing the proposal distribution, thus making the algorithm even more efficient.

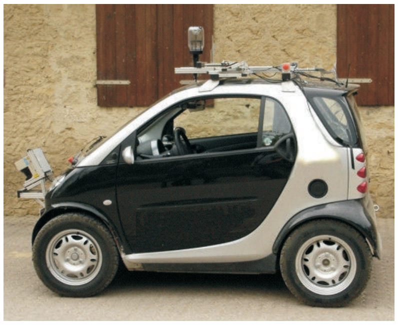


Fig. 5. Our vehicle equipped with an omnidirectional camera.

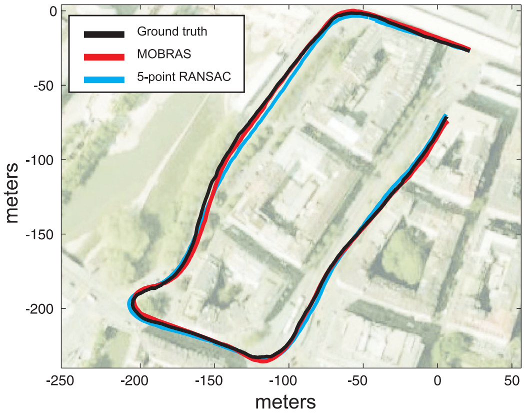


Fig. 6. Comparison between estimated trajectories. (black) Ground truth. (red) MOBRAS. (cyan) 5-point RANSAC.

has the problem that the estimated trajectory is affected by drift. The drift appears either as an effect of the motion error accumulated over time or as the consequence of spurious errors in the motion-estimation algorithm, which can cause sudden deviations of the estimated trajectory. With the second criterion, conversely, spurious errors can be detected very easily.

B. Trajectory estimation

The comparison between the trajectories estimated by the two algorithms is shown in figures 6 and 7 for two different paths. We fixed the number of iterations of the 5-point RANSAC to 1000. This number was calculated using expression (1) assuming a 99% confidence and 50% of outliers in the data. According to (1) the minimum number of iteration of 5-point RANSAC should be 145. However, it is common by multiply this factor by 10 to increase the probability that the returned solution is the one with the largest consensus. Conversely, we tested MOBRAS with both 100 and 1000 samples. The two trajectories shown in these figures were both obtained with just 100 samples. As can be observed, in both cases the path estimated by MOBRAS is closer to the ground truth than that estimated by the 5-point RANSAC. This result confirms what we discussed in the preceding sections, that is, MOBRAS is expected to be more accurate than RANSAC because it *captures* the dominant component of the motion.

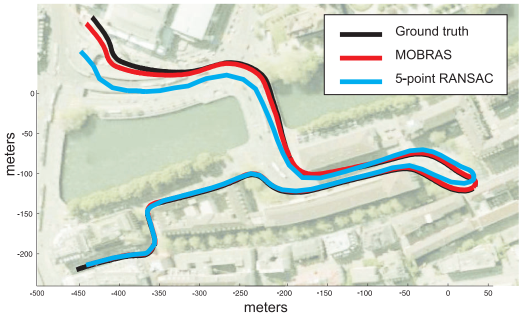


Fig. 7. Comparison between estimated trajectories. (black) Ground truth. (red) MOBRAS. (cyan) 5-point RANSAC.

C. Orientation error

In Fig. 8, we compare the error in orientation between the two algorithms with respect to the ground truth for the trajectory in Fig. 7. The error is calculated for the roll, pitch, and yaw angles. First of all, we can notice that the errors of the two algorithms are very well correlated. Additionally, they look very similar. However, notice that the yaw error introduced by the 5-point RANSAC was larger than that of MOBRAS in four occasions. After visual inspection, we found that this was not due to a lower number of inliers found by RANSAC, but rather to different inliers.

D. Execution time

Finally, we compared the execution time of the two algorithms for the case of 100 samples (MOBRAS) and 100 iterations (5-point RANSAC).[13] The distribution of the execution times is shown in Fig. 9. It can be noticed that MOBRAS is about 50 times faster than 5-point RANSAC.[14] This result is perfectly in agreement with the discussion in Section V-B.

VII. Conclusions

In this paper, we presented an algorithm (MOBRAS) to compute both the relative motion of a single camera and wrong data associations, which does not rely on a RANSAC-scheme.

Visual odometry was posed as a maximum-likelihood estimation problem. The algorithm operates by estimating a proposal distribution that *captures* the main components of the motion, and which is based on the vehicle non-holonomic constraints. We showed that sampling from this proposal is equivalent to computing the joint posterior probability of the complete motion. Alternatively, if one is looking only for the maximum-likelihood solution, then one can choose the solution that has given more inliers.

We successfully tested the algorithm on a large image dataset collected from a car while driving in a urban environment. The results of MOBRAS were compared against

[13] We used the implementation of the 5-point algorithm available from the authors' webpage [20].

[14] The execution time of both algorithms scales linearly with the number of samples/iterations.

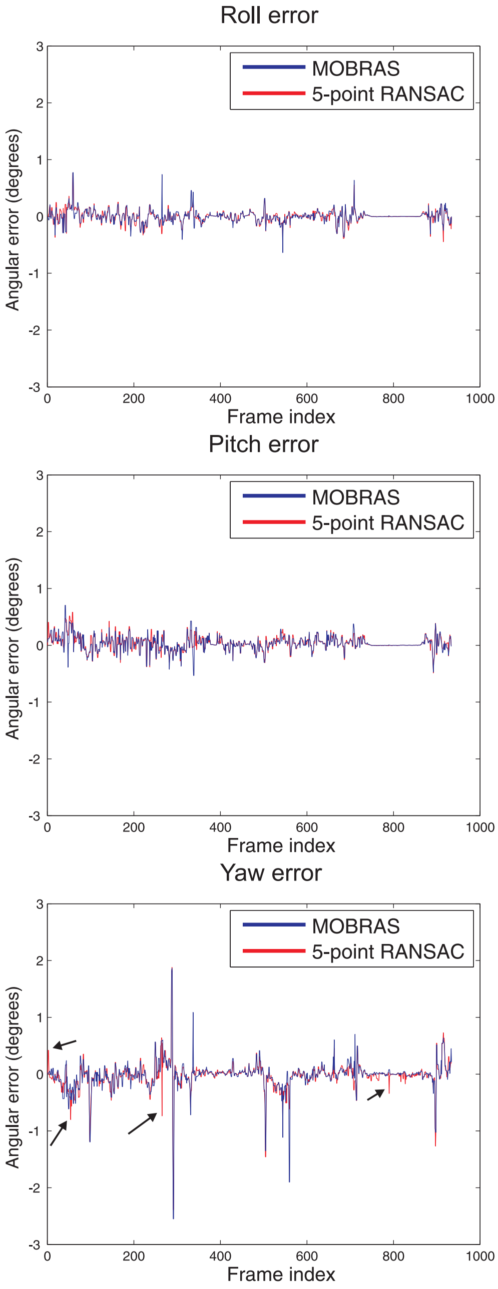


Fig. 8. Comparison between estimated orientations. (blue) MOBRAS. (red) 5-point RANSAC. The arrows indicate spots where the error of the 5-point RANSAC was larger than that of MOBRAS. This happened, however, only four times over 1000 frames.

those of the well known 5-point RANSAC using ground-truth data. We showed that MOBRAS is significantly faster (by a factor of 50) than the 5-point RANSAC. Additionally, we showed that the returned solution is more accurate in that it favors the underlying motion model of the vehicle, thus overcoming the typical limitations of RANSAC.

Finally, the proposed algorithm provides the likelihood of the motion estimate, which can be very useful in all those applications where a probability distribution of the position of the vehicle is required (e.g., SLAM).

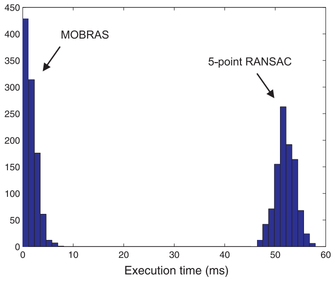


Fig. 9. Comparison between execution times.

References

[1] D. Nister, O. Naroditsky, , and B. J., "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, no. 1, 2006.

[2] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers: Field reports," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.

[3] R. Goecke, A. Asthana, N. Pettersson, and L. Petersson, "Visual vehicle egomotion estimation using the fourier-mellin transform," in *IEEE Intelligent Vehicles Symposium*, 2007.

[4] J. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'08)*, 2008.

[5] M. J. Milford and G. Wyeth, "Single camera vision-only slam on a suburban road network," in *IEEE International Conference on Robotics and Automation (ICRA'08)*, 2008.

[6] D. Scaramuzza and R. Siegwart, "Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles," *IEEE Transactions on Robotics, Special Issue on Visual SLAM*, vol. 24, no. 5, October 2008.

[7] M. A. Fischler and R. C. Bolles, "RANSAC random sampling concensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of ACM*, vol. 26, pp. 381–395, 1981.

[8] D. Nistér, "An efficient solution to the five-point relative pose problem," in *CVPR03*, 2003, pp. II: 195–202.

[9] O. Chum and J. Matas, "Matching with prosac - progressve sample consensus," in *CVPR*, 2005.

[10] D. Nister, "Preemptive ransac for live structure and motion estimation," *Machine Vision and Applications*, vol. 16, no. 5, pp. 321–329, 2005.

[11] R. Raguram, J. Frahm, and M. Pollefeys, "Exploiting uncertainty in random sample consensus," in *ICCV*, 2009.

[12] J. Civera, O. Grasa, A. Davison, and J. Montiel, "1-point ransac for ekf filtering: Application to real-time structure from motion and visual odometry," *Journal of Field Robotics*, vol. 27, pp. 609–631, 2010.

[13] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *IEEE International Conference on Robotics and Automation (ICRA'09)*, 2009.

[14] D. Scaramuzza, "1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *International Journal of Computer Vision*, vol. 95, no. 1, 2011.

[15] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[16] B. Micusik and T. Pajdla, "Omnidirectional camera model and epipolar geometry estimation by ransac with bucketing," in *Scandinavian Conference on Image Analysis (SCIA)*, 2003.

[17] R. Siegwart, I. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots, second edition*. MIT Press, 2011.

[18] P. Lamon, S. Kolski, and R. Siegwart, "The smartter - a vehicle for fully autonomous navigation and mapping in outdoor environments," in *CLAWAR*, 2006.

[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[20] H. Stewenius, C. Engels, and D. Nister, "Recent developments on direct relative orientation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 60, pp. 284–294, 2006.