REAL-TIME SLAM FOR THE OFF-ROAD AUTONOMOUS DRIVING

B. Vishnyakov *, V. Sheverdin

FGUP «State Research Institute of Aviation Systems», Russia, 125319, Moscow, Viktorenko street, 7
(vishnyakov, sheverdin)@gosniias.ru

Commission II, WG II/4

KEY WORDS: SLAM, off-road, autonomous driving, DAS, scene reconstruction, SfM, odometry.

ABSTRACT:

In this paper we propose a new SLAM algorithm that is robust to the changing environment of the countryside. The hardware part consists of two separate machine vision cameras, joined in stereo, and can be supplemented with LiDAR, IMU and GPS. We introduce a method that can be used to reliably calculate the position of a vehicle in natural environments. To estimate the pose and produce a three-dimensional reconstruction we use a stereo camera rig, inertial measurement unit and the global positioning system. While solving the problem of visual odometry in outdoor scenes we faced a number of difficulties, arising from high dynamic range, as well as the presence of a large number of “similar elements”, such as leaves, grass, trees and etc. Under these conditions, it becomes difficult to match feature points in image sequences. HOG-based methods, such as SIFT, SURF and others often do not obtain good matching due to noise, lack of a sufficient number of gradients, and the presence of identical domains. Using neighborhood-based detectors such as DAISY often allows to identify the correct matches, but using them is worth it too expensive. These methods are very demanding on the computational resources and prone to drift. We needed a method that is less expensive, but at the same time provides sufficient accuracy in the trajectory estimation. Direct methods, such as optical flow calculating or direct image matching allow us to map point-to-point in these conditions with high reliability. They also have disadvantages that can be eliminated by using an IMU and modern algorithms. To improve the quality of the algorithms, we solve the reconstruction problem for several frames using the Levenberg-Marquardt optimization method for bundle adjustment. Each pass optimizes frames that are directly related to the last one, we use two threads that perform partial and full optimization of the entire trajectory using graphs to significantly increase the performance of the method.

1. INTRODUCTION

Methods of simultaneous localization and mapping use sensors of a very different nature to improve their quality. In particular, visual sensors, LiDARs, inertial measuring devices, GLONASS and GPS systems, and radars are used. In this paper, we consider using two machine vision cameras combined into stereo, IMU and GPS to solve the problem of SLAM. Simultaneous localization and three-dimensional scene reconstruction involve the joint solution of two subtasks: determining own position (odometry) and reconstructing a three-dimensional scene (building a spatial model).

There are basically two classes of methods for solving the SLAM problem using visual sensors: feature-based and direct.

The first one is based on matching feature-based image tracking and performs sparse reconstruction of three-dimensional space (Sparse reconstruction). The second set of methods, called direct methods, compares the images themselves or some of their parts connected by topology, allowing you to get a dense and semi-dense reconstruction of space (Dense and Semi-dense reconstruction).

Feature-based class of methods includes such implementations as PTAM (Klein, Murray, 2007), monoSLAM (Daivison et al., 2007), ORB-SLAM (Mur-Artal, Tardós, 2017), ProSLAM (Schlegel et al., 2018), OpenVSLAM (Sumikura et al., 2019) e.g. in Figure 1. Direct class includes, for example, DSO (Engel et al., 2018), SVO (Forster et al., 2014), DTAM (Newcombe et al., 2011), LSD-SLAM (Engel, 2017) can be seen in Figure 2.

Figure 1. Sample of ORB-SLAM.

Figure 2. Sample of LSD-SLAM.

These approaches have both advantages and disadvantages. For example, direct methods show good stability, although they require some pre-processing. They require fewer computing resources, but at the moment the class of these odometry methods is still undergoing a stage of active development and has some problems. Good results can be achieved by combining the solution of direct and indirect methods using the Kalman filter and other Bayesian filters (Lee et al., 2019).

* Corresponding author

This contribution has been peer-reviewed.

Feature point detection and matching techniques, based on histograms of oriented gradients, such as SIFT (Lowe, 2004), SURF (Bay et al., 2008) and others often do not obtain good matching due to noise and insufficient number of gradients, presence of identical domains. Neighborhood-based feature point detectors such as DAISY (Tola et al., 2008) often allow to identify the correct matches, but using them reduces the algorithm speed drastically.

Reliable solution of the SLAM problem requires joint modelling using a variety of methods, sensors, and probabilistic filtering algorithms – sensor fusion. Moreover, the science society is now actively researching algorithms that work with event cameras, which are able to increase the stability of reconstruction in conditions of fast movement and insufficient lighting.

In this paper we propose a complex solution for off-road and countryside scene reconstruction. Our method is based on the following key steps:

- Gradient-based camera auto exposure.
- Adaptive filtering of captured images to reduce noise.
- Fast stereo correspondence calculation
- Using IMU measurements for sensor initial extrinsic
- Using a semi-direct method for calculating visual odometry
- Obtaining semi-dense reconstruction by using DAISY features descriptor or block matching algorithm

2. EXPOSURE CORRECTION AND FILTERING

During full-scale tests we almost every time encountered overexposed and underexposed regions within a single outdoor image due to high dynamic range. To overcome this issue, we use auto exposure control method for maximizing the number of features in the image (Zang et al., 2017). The more differences (gradients) of brightness can be distinguished, the more accurate and high-quality the result will be achieved.

There is a large number of methods for correcting the effects of brightness changes, ranging from heuristic to methods, based on neural networks and machine learning. High quality results are obtained using the camera response function. This is a function linking the pixel brightness and the natural logarithm of the exposure time of a frame. We correct the exposure time based on the camera response function on Figure 4 (Debevec, Malik, 2008).

If there is not enough information about gradients, the standard auto-exposure method, based on the brightness distribution histogram, is used.

![Figure 4. Camera response function.](image)

After capturing images, adaptive filtering is performed to suppress noise in areas with small gradients. Just as in the previous step, the image is filtered with a median filter with a 3-pixel frame window, and both images are combined by the gradient function:

\[ I = f(|\nabla I_s|) \ast I_s + (1 - f(|\nabla I_s|)) \ast I_f \]

Here \( I \) is the resulting pixel intensity, \( f \) is the weight function of the gradient intensity, \( I_s \) is the original image, and \( I_f \) is the filtered value intensity. Thus, in areas with large gradients, the original values are used, while in places with small gradients, they are filtered out (Figure 5).

![Figure 5. Adaptive noise suppression](image)
3. STERERO CORRESPONDENCIES

There are many methods for pixel matching in stereo normal case, from simple block matching to convolution neural network approaches.

A point is called a “point of interest” or an “interest point” (Figure 6) if there are a sufficient number of characteristics in its neighbourhood that make it possible to distinguish it from others and compare with points in another image.

![Figure 6. Interest Points.](image)

In frames that contain a large number of similar elements, the relative position of the interest points must be calculated.

Given two images $I_1$ and $I_2$ in the sequence, we can formulate the following statement:

$$l_2 = l_1(x + u(x, y), y + v(x, y))$$  \hspace{1cm} (3)

where $u(x, y), v(x, y)$ denote the displacement of pixel between images. In the stereo normal case, we can write:

$$u = \Delta x$$

$$\Delta x \geq 0$$  \hspace{1cm} (4)

$$v = 0$$

$$l_2 = l_1(x + \Delta x, y)$$

In the linear case, for infinitesimal $u, v$:

$$l_2 = l_1 + l_{1x}u + l_{1y}v$$  \hspace{1cm} (5)

or, in stereo normal case:

$$l_2 = l_1 + l_{1x}dx$$  \hspace{1cm} (6)

After that, the problem is solved by using the standard pyramid method of optical flow (Lucas, Kanade, 1981). Using this approach allows us to significantly speed up the comparison process, so we can quickly establish the stereo correspondences (Figure 7).

![Figure 7. Stereo correspondences of some points.](image)

To achieve maximum efficiency and speed of the algorithm, we limit the number of considered interest points to the minimum value, necessary for a reliable solution, by iteratively adding the new elements and discarding the unreliable ones until an acceptable accuracy is achieved. We believe that to reliably determine the position of the camera at the next frame, it is necessary to correlate $K = 2^7$ elements located throughout the entire image. First, we select and compare $N = 2^8$ most prominent singular points in different parts of the image. We discard unreliable items through RANSAC fundamental matrix estimation algorithm (Fischler, Bolles, 1981). In case of insufficient number of matched interest points, the initial number of points doubles, and the procedure is repeated. This approach allows us to increase the speed of the algorithm by using only a small number of features in each frame.

4. VISUAL-INERTIAL ODOMETRY

The next step is evaluation of our own trajectory. First, we perform triangulation of the corresponding points of the stereo rig and determine the mathematical expectation and dispersion of their spatial coordinates $(\mu_x, \mu_y, \mu_z, \sigma_x, \sigma_y, \sigma_z)$.

A sufficient distance between the sensors allows us to determine the spatial coordinates with great accuracy, but at the same time it increases the frequency of mapping errors due to distortions and occlusions. Each spatial point in addition to mathematical expectation and variance also contains the baseline length value for determining outlier probability. A shorter baseline length is less likely to match incorrectly, and can be used by subsequent RANSAC algorithm.

After the stereo reconstruction, we need to calculate the trajectory of the camera when moving to the next frame. The direct method (3) is used for this purpose. The functions $u$ and $v$ (3) are defined by a projective transformation of consecutive frame in homogeneous coordinates:

$$(u, v, \lambda) = K [R_2 \begin{bmatrix} l_1 \end{bmatrix} - l_1 \begin{bmatrix} 0 \end{bmatrix}]$$  \hspace{1cm} (7)

Here $K$ is the camera matrix, $R_2, X_2$ specify the relative movement of the camera, $p$ is the point of the three-dimensional space, and $\lambda$ is the homogeneous coefficient. Since these relations are in homogeneous coordinates, the equations for $u$ and $v$ are not linear and contain irrationality.

Minimization error cannot be performed using the least squares approach by SVD decomposition, so iterative algorithms, such as Levenberg-Marquadt should be applied. For some of the equations, the coordinates of point $p$ are known from the stereo reconstruction; for others we need to select several close points to obtain the desired number of equations. We assume that very close points have the same spatial coordinates.
Iterative optimization algorithms require a good initial approximation, for this purpose we use the IMU data. Also, these algorithms suffer from outliers, which is why we use the RANSAC algorithm. We also use the g2o library to implement windowed Bundle Adjustment in separate thread (Kuemmerle et al., 2011). For solving optimization task, we construct a dynamic structure – a graph (Figure 9), which describes the relative position of the sensors, points of three-dimensional space, their projections at different points in time and covariance matrix.

After solving the visual odometry equations, we combine the results with the IMU and GPS data using the Extended Kalman Filter approach (Thrun et al., 2005). The resulting trajectory is shown in Figure 10.

5. SEMI-DENSE RECONSTRUCTION

There are different approaches for stereo mapping, such as block matching, using global descriptors, and dense optical flow, deep learning and others. In our work, we use DAISY descriptors and a simple block matching method. A semi-dense point cloud is calculated twice per second and aligned in the world frame. This allows us to perform a good quality reconstruction in real time (Figure 11).

Figure 8. Points tracing

Figure 9. Scene graph.

Figure 10. Sparse visual-inertial odometry

ACKNOWLEDGEMENTS

The reported study was funded by RFBR project № 19-07-01248 A.

REFERENCES


