Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-10 (2015) 227–232

Monocular Visual Odometry on a Smartphone

Simon Tomažič*, Igor Škrjanc*

**Faculty of Electrical Engineering, Laboratory of Autonomous Mobile Systems, University of Ljubljana, Ljubljana, Slovenia, (e-mail: simon.tomazic@fe.uni-lj.si, igor.skrjanc@fe.uni-lj.si)}*

Abstract: The paper presents the monocular visual odometry, which is the sequential estimation process of camera motions depending on the perceived movements of pixels in the image sequence. The visual odometry accurately determines the incremental movements and the positions of the camera according to the world coordinate system. The algorithm consists of four other algorithms, namely camera calibration algorithm, KLT algorithm, algorithm for estimation of rigid transformation and RANSAC algorithm. Since the visual odometry algorithm was developed with the purpose to be a part of an indoor localization application for pedestrians and especially the blind, it was fully implemented on a smartphone.

Keywords: Calibration, KLT, Rigid transformation, RANSAC, Visual odometry, Smartphone

1. INTRODUCTION

All living creatures on Earth have, to a certain extent, developed the ability to determine their position in an environment. With the development of new localization methods and algorithms, it is desired to enable robots, autonomous mobile systems, and particularly people who have lost this ability, e.g. due to blindness, to use this ability.

In our studies the focus is on indoor localization, since this represents a challenge, mostly due to the fact that GPS signals are not available there. As smartphones have become essential accessories of a modern man, the possibility of using their hardware for the purpose of indoor localization of autonomous mobile systems and persons has been studied. Modern smartphones include multiple sensors such as the gyroscope accelerometer, altimeter, magnetometer, camera and multiple communication modules (WiFi, Bluetooth, LTE, NFC), which can be used together with different algorithms for indoor localization.

Currently the most widely used approach to indoor localization using a smartphone is based on the measurement of the WiFi, Bluetooth or GSM signal strength. However, because this approach does not enable high positioning accuracy, new approaches based on the fusion of several different sensors and methods are being established. In particular by using the Kalman filter it is possible to effectively combine the information from the camera and inertial sensors.

In the paper, the focus is on using the camera as the sensor for precise estimation of the incremental movements of a smartphone, thereby a pedestrian (or a blind person) wearing a smartphone. The algorithm described below represents only a part of the final application which could help the blind to localize themselves in an indoor environment and which could inform them about the obstacles in their path.

In the field of visual localization several different methods and algorithms which can determine the movement of an agent (a vehicle, person, robot) equipped with a camera have already been developed. Among the established methods, here belong Simultaneous Localization and Mapping (SLAM), Structure from Motion (SFM) and Visual Odometry (VO). SLAM and SFM require powerful CPU and quite a lot of memory, because they build a 3D map of the environment besides the motion estimation. On the other hand, VO estimates only the motion of the camera and consequently it can operate in real time, even on less powerful hardware. The concept of visual odometry was established by Nister et al. (2004), who introduced the main concept, which is the basis for the most of the existing VO algorithms. Our algorithm is based on the assumption that the smartphone is fixed on a certain height and at an angle relative to the floor. Accurate transformation between the camera C. S. and the floor C. S. is obtained by initial calibration. In relation to the visual odometry, RANSAC algorithm (see Nister et al. (2004)) is often used, which allows the elimination of outliers in the calculation of the rigid motion model. For the purpose of testing the VO algorithm, a Galaxy S4 smartphone based on the operating system Android was used. In the implementation of the application, an open-source library BoofCV that is written in the Java was used. In the application the images were captured at a resolution of 320×240 pixels. The calibration was also performed on the smartphone by using the BoofCV library (see BoofCV (2014)).

In the following sections the components of the monocular VO are first presented, then VO itself and finally the experiment results of the implemented algorithm functionality.

2. MONOCULAR VISUAL ODOMETRY

Monocular visual odometry sequentially estimates camera motions according to the perceived movements of pixels in the image sequence. The visual odometry algorithm is composed of four algorithms, namely: the camera calibration, the feature tracker, the algorithm for the estimation of a rigid motion model and the RANSAC algorithm.

2.1 Camera calibration

The camera for which it is assumed that it has a thin lens, can be described with the pinhole camera model (see Sonka et al.

10.1016/j.ifacol.2015.08.136

(2014)). Let a point located in the image be denoted with $\boldsymbol{m} = [u, v]^T$ and a point located in the 3D space with $\boldsymbol{M} = [X, Y, Z]^T$ (see Zhang (2000)). An optical ray that is reflected from a point $\boldsymbol{M}$ of the observed scene and travels through the optical center C determines a projected point $\boldsymbol{m}$ in the image plane. Points $\boldsymbol{m}$ and $\boldsymbol{M}$ can be expressed in homogeneous coordinates as $\boldsymbol{m} = [u, v, 1]^T$ and $\boldsymbol{M} = [X, Y, Z, 1]^T$. The transformation between the 3D point $\boldsymbol{M}$ and its projection on the image - point $\boldsymbol{m}$ is defined by the pinhole camera model:

$$s\boldsymbol{m} = \boldsymbol{A}[\boldsymbol{R}\ \boldsymbol{T}]\boldsymbol{M} \tag{1}$$

where s is a scale factor. Rotation matrix $\boldsymbol{R}$ and translational vector $\boldsymbol{T}$ which describe the transformation between the world coordinate system and the camera coordinate system represent the extrinsic parameters. $\boldsymbol{A}$ denotes a matrix of camera intrinsic parameters which is defined as:

$$\boldsymbol{A} = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where $\alpha = fs_x$ and $\beta = fs_y$ represent the focal length expressed in pixels according to the coordinate axes u and v. f is the focal length expressed in millimeters, s_x and s_y are scale factors according to the coordinate axes u and v, which determine number of pixels per millimeter and $(u_0\ v_0)$ is the principal point. Skew parameter γ represents the distortion of a pixel.

As the most appropriate mode for the estimation of the planar homography between calibration target (the plane is covered with a chessboard pattern) and its image, a method that is based on the criterion of maximum likelihood was chosen (see Zhang (2000)). In this approach the calibration parameters are estimated analytically in the first step and in the second step this result is optimized by using the non-linear optimization technique based on the maximum likelihood criterion.

In the case when the calibration target is used, it may be assumed that Z component of the point $\boldsymbol{M}$ is always equal to zero. For this reason the point $\boldsymbol{M}$ can be written as $\boldsymbol{M} = [X\ Y]^T$ or in homogeneous coordinates as $\boldsymbol{M} = [X\ Y\ 1]^T$. Consequently the transformation between points $\boldsymbol{m}$ and $\boldsymbol{M}$ can be expressed as:

$$s\boldsymbol{m} = \boldsymbol{H}\boldsymbol{M} \tag{3}$$

where $\boldsymbol{H} = \boldsymbol{A}[\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{T}]$ is homography defined up to scale factor λ. Further the next relation can be defined:

$$[\boldsymbol{h}_1\ \boldsymbol{h}_2\ \boldsymbol{h}_3] = \lambda\boldsymbol{A}[\boldsymbol{r}_1\ \boldsymbol{r}_2\ \boldsymbol{T}], \tag{4}$$

where h_i is i -th column of the matrix $\boldsymbol{H}$ and r_i is i -th column of the matrix $\boldsymbol{R}$. The maximum likelihood estimation of homography $\boldsymbol{H}$ is gained by minimizing the expression:

$$\textstyle\sum_i(\boldsymbol{m}_i - \widehat{\boldsymbol{m}}_i)^T\boldsymbol{\Lambda}_{\boldsymbol{m}_i}^{-1}(\boldsymbol{m}_i - \widehat{\boldsymbol{m}}_i) \tag{5}$$

where $\widehat{\boldsymbol{m}}_i = \frac{1}{\bar{\boldsymbol{h}}_3^T\boldsymbol{M}_i}\begin{bmatrix}\bar{h}_1^T M_i \\ \bar{h}_2^T M_i\end{bmatrix}$ is the point obtained from the model of the calibration target points, $\boldsymbol{\Lambda}_{\boldsymbol{m}_i} = \sigma^2\boldsymbol{I}$ is the covariance matrix and $\bar{\boldsymbol{h}}_i$ is i -th row of the matrix $\boldsymbol{H}$. The equation (5) can be further presented as a nonlinear minimization problem, which can be solved with the method of least squares $min_H \sum_i\|\boldsymbol{m}_i - \widehat{\boldsymbol{m}}_i\|^2$. To get the analytic solution, all rows of the matrix $\boldsymbol{H}$ are further written in matrix $x = [\bar{\boldsymbol{h}}_1^T, \bar{\boldsymbol{h}}_2^T, \bar{\boldsymbol{h}}_3^T]^T$ and the equation (3) is transformed in:

$$\begin{bmatrix} \boldsymbol{M}^T & 0^T & -u\boldsymbol{M}^T \\ 0^T & \boldsymbol{M}^T & -v\boldsymbol{M}^T \end{bmatrix}\boldsymbol{x} = 0 \tag{6}$$

or $\boldsymbol{Lx} = 0$, where $\boldsymbol{L}$ is a matrix with dimensions $2n$×9 (n is the number of points). As the matrix $\boldsymbol{x}$ is defined up to scale factor, the solution is the right singular vector of the matrix $\boldsymbol{L}$, which is associated with the smallest singular value.

Let symmetric matrix $\boldsymbol{B} = \boldsymbol{A}^{-T}\boldsymbol{A}^{-1}$ be defined as 6D vector $\boldsymbol{b} = [B_{11}, B_{12}, B_{22}, B_{13}, B_{23}, B_{33}]^T$.

If $\boldsymbol{h}_i = [h_{i1}, h_{i2}, h_{i3}]^T$ is the i–th column of the matrix $\boldsymbol{H}$, then the next equation can be written as:

$$\boldsymbol{h}_i^T\boldsymbol{B}\boldsymbol{h}_j = \boldsymbol{v}_{ij}^T\boldsymbol{b} \tag{7}$$

where $\boldsymbol{v}_{ij} = [\boldsymbol{h}_{i1}\boldsymbol{h}_{j1},\ \boldsymbol{h}_{i1}\boldsymbol{h}_{j2} + \boldsymbol{h}_{i2}\boldsymbol{h}_{j1},\ \boldsymbol{h}_{i2}\boldsymbol{h}_{j2},\ \boldsymbol{h}_{i3}\boldsymbol{h}_{j1} + \boldsymbol{h}_{i1}\boldsymbol{h}_{j3},\ \boldsymbol{h}_{i3}\boldsymbol{h}_{j2} + \boldsymbol{h}_{i2}\boldsymbol{h}_{j3},\ \boldsymbol{h}_{i3}\boldsymbol{h}_{j3}]^T$.

With taking into consideration that the vectors $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ are orthonormal, the equation for the limitation of intrinsic parameters can be given as:

$$\begin{bmatrix} \boldsymbol{v}_{12}^T \\ (\boldsymbol{v}_{11} - \boldsymbol{v}_{22})^T \end{bmatrix}\boldsymbol{b} = 0 \tag{8}$$

For n captured images of the calibration target also n equations (8) can be written. These can be further combined in a new equation as:

$$\boldsymbol{Vb} = 0 \tag{9}$$

where $\boldsymbol{V}$ is matrix with the size of $2n$×6. To get a unique solution for vector $\boldsymbol{b}$, the number n has to be $n \geq 3$. The equation (9) can be solved by using the singular value decomposition (SVD), where the vector $\boldsymbol{b}$ is equal to the eigenvector of matrix $\boldsymbol{V}^T\boldsymbol{V}$, which belongs to the smallest eigenvalue.

When the vector $\boldsymbol{b} = [B_{11}, B_{12}, B_{22}, B_{13}, B_{23}, B_{33}]^T$ is estimated, all intrinsic parameters can be computed as:

$$\begin{gathered} v_0 = (B_{12}B_{13} - B_{11}B_{23})/(B_{11}B_{22} - B_{12}^2) \\ \lambda = B_{33} - (B_{13}^2 + v_0(B_{12}B_{13} - B_{11}B_{23}))/B_{11} \\ \alpha = \sqrt{\lambda/B_{11}}, \beta = \sqrt{\frac{\lambda B_{11}}{B_{11}B_{22}-B_{12}^2}} \\ \gamma = -B_{12}\alpha^2\beta/\lambda, u_0 = \frac{\gamma v_0}{\beta} - B_{13}\alpha^2/\lambda \end{gathered} \tag{10}$$

When the matrix of the intrinsic parameters $\boldsymbol{A}$ is computed, also extrinsic parameters can be obtained by taking into account the equation (3):

$$\boldsymbol{r}_1 = \lambda\boldsymbol{A}^{-1}\boldsymbol{h}_1, \boldsymbol{r}_2 = \lambda\boldsymbol{A}^{-1}\boldsymbol{h}_2, \boldsymbol{r}_3 = \boldsymbol{r}_1 \times \boldsymbol{r}_2, \boldsymbol{T} = \lambda\boldsymbol{A}^{-1}\boldsymbol{h}_3 \tag{11}$$

where $\lambda = \frac{1}{\|A^{-1}h_1\|} = \frac{1}{\|A^{-1}h_2\|}$. In the calibration process the radial distortion described by the following model was also considered:

$$\begin{bmatrix}\acute{x} \\ \acute{y}\end{bmatrix} = (1 + k_1r^2 + k_2r^4)\begin{bmatrix}x \\ y\end{bmatrix} \tag{12}$$