

Fusion of IMU and Vision for Absolute Scale Estimation in Monocular SLAM

ENPM667 Midterm Project Report

Harsh Kakashaniya · Rohitkrishna Nambiar

Submitted: 19th November 2018

Abstract The fusion of inertial and visual data is widely used to improve an objects pose estimation. However, this type of fusion is rarely used to estimate further unknowns in the visual framework. In this paper we present and compare two different approaches to estimate the unknown scale parameter in a monocular SLAM framework. Directly linked to the scale is the estimation of the objects absolute velocity and position in 3D. The first approach is a spline fitting task adapted from Jung and Taylor and the second is an extended Kalman filter. Both methods have been simulated offline on arbitrary camera paths to analyze their behavior and the quality of the resulting scale estimation. We then embedded an online multi rate extended Kalman filter in the Parallel Tracking and Mapping (PTAM) algorithm of Klein and Murray together with an inertial sensor. In this inertial/monocular SLAM framework, we show a real time, robust and fast converging scale estimation. Our approach does not depend on known patterns in the vision part nor a complex temporal synchronization between the visual and inertial sensor.

Keywords IMU vision fusion · Absolute scale · Monocular SLAM · Kalman filter

Harsh Kakashaniya
UID: 116311236
E-mail: harshbk@umd.edu

Rohitkrishna Nambiar
UID: 115507944
E-mail: rohit517@umd.edu

Contents

1	Introduction	4
2	Hardware Setup	4
3	Camera and image formation process	4
4	What's wrong with monocular camera?	6
5	Spline Fitting Method	7
5.1	Types of curve fitting	8
5.1.1	Maximum error	8
5.1.2	Average error	8
5.1.3	RMS (Least square method)	8
5.2	IMU Reading	9
5.3	Results	12
6	Extended Kalman Filter	14
6.1	Introduction	14
7	Simultaneous Localization and Mapping	16
7.1	Visual Odometry	17
7.2	MonoSLAM	19
7.3	Parallel Tracking and Mapping (PTAM)	19
7.4	Results	21
7.4.1	Simulated and Real Data	21
7.4.2	Online Implementation	22

List of Figures

1	a. Camera/IMU setup. b. Frame transformations	5
2	Please write your figure caption here	5
3	Please write your figure caption here	6
4	Monocular camera image coordinates	7
5	Curve Fitting with Maximum error	9
6	Curve Fitting with Average error	9
7	Curve Fitting with RMS error	10
8	In this case Blue is actual curve traced by IMU this are points with noise.And green is plotted curve. With least square method.	12
9	Multiple spline fit	13
10	A. Red curve for method with one curve and least square method. B. Green curve is using the given method by Jung and Taylor.	14
11	SLAM architecture	17
12	Generalized Visual Odometry Pipeline	18
13	Offline simulation results with simulated and real data. Fewer directions (X,Y,Z) included in Kalman filter gives us a better estimate of λ . [11]	21
14	Online implementation with PTAM work flow.Blue color boxes are original PTAM threads. Yellow boxes are added threads for scale estimation λ . [11]	22

1 Introduction

On-board pose estimation is very useful for a variety of applications ranging from autonomous robotics to augmented reality. Pose estimation can be done using different sensors such as cameras and inertial measurement units (IMU). Simultaneous localization and mapping (SLAM) using a monocular camera can be used to get the trajectory of the camera and a map of the environment. But monocular cameras suffer from scale ambiguity causing the overall trajectory to drift making it unusable in real time. Using a stereo camera helps solve the scale ambiguity. IMUs on the other hand can be used to get the trajectory traveled by integrating the acceleration measurements over time. But this leads to highly inaccurate trajectory estimates. The estimate of the scale factor is essential to fuse both these measurements. This fusion helps us determine the unknown scale factor λ .

In this paper [11], two methods are presented for scale estimation. The first one is a spline fitting method by Jung and Taylor [4]. The second is a multi rate Extended Kalman Filter(EKF). Both the approach have been simulated in MATLAB.

This report is organized as follows. Section 2 goes over the camera and imu setup, Section 3 covers the camera and image formation process, Section 4 goes over the scale estimation problem with monocular camera, Section 5 outlines the spline fitting method with results.

2 Hardware Setup

This section goes through the hardware setup used in the research paper. USB uEye UI-122xLE fisheye camera was used as the vision input. The camera has a resolution of 752 480 and a frame rate up to 87 fps. Motion blur was minimized due to the high dynamic range and global shutter of the camera. VG400CC-200 solid state gyro which includes a tri-axial gyroscope and tri-axial accelerometer was used. It has an output frequency of 75 Hz with an input range of $\pm 10g$ and ± 1.25 mg resolution. The IMU outputs acceleration around its 3-axis along with yaw, pitch and roll.

3 Camera and image formation process

To understand how a camera perceives the environment, it is necessary we understand the image formation process. For us humans, we see things when a light originating from a light source is reflected on an object and enters our eyes. A camera acts very much similar to the human eye. The earliest and first model of an optical camera is the pinhole camera which is a simple and highly accurate representation of our eye model. This is the simplest device to form an image of a 3D scene on a 2D surface. As seen in figure 1 rays of light enter the pinhole and forms an inverted image of the object. This is called

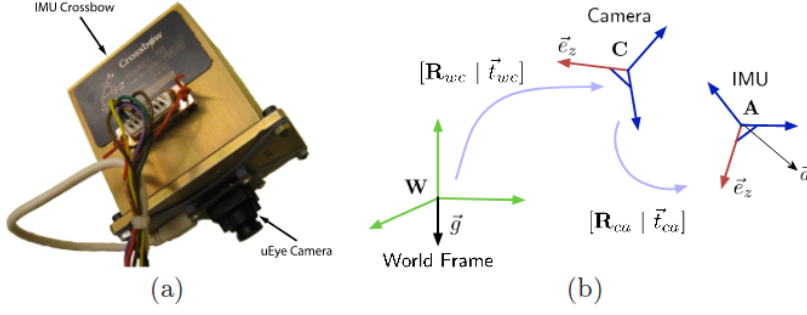


Fig. 1 a. Camera/IMU setup. b. Frame transformations

perspective projection. As the image formed in the image plane is inverted, we consider a virtual plane in front of the pinhole that acts as the image plane.

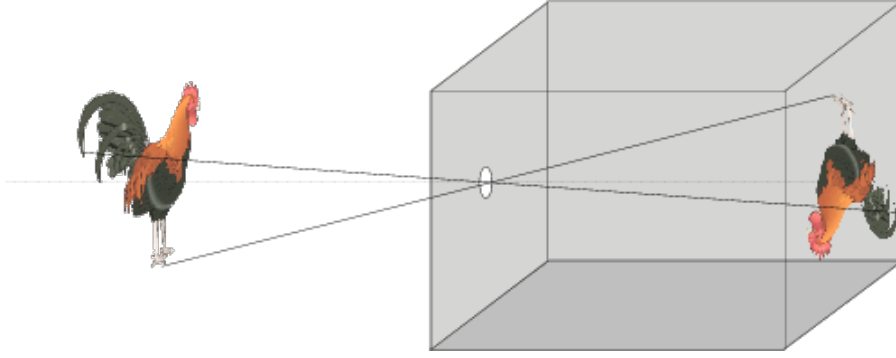


Fig. 2 Please write your figure caption here

Although the pinhole model is quite accurate, modern cameras have lenses. They help gather more light from a source leading to higher quality sharper images. Ignoring the internal diffraction, we assume thin lens equations for the imaging process. Every camera has a region of depths over which the scene is sharp. Modern cameras have variable apertures which help in focusing on objects at varying distances in the scene.

Every camera has an intrinsic, extrinsic and distortion parameters that are important to understand for every application. These parameters are used to correct for lens distortion, measure objects in physical world and to determine the location of the camera in the real world. Camera calibration is the process of estimating these parameters specific to a given camera and application setup. Figure 2 demonstrates how the a real world 3D coordinate is related to a 2D pixel coordinate using the parameters mentioned above.

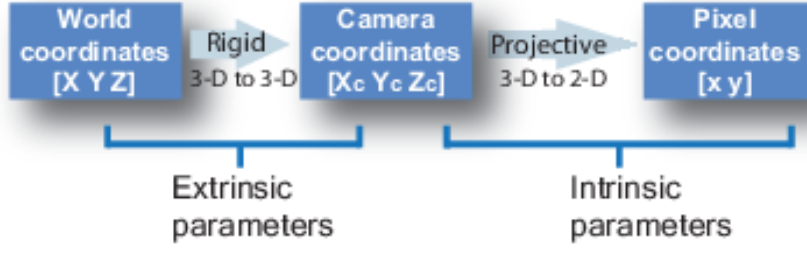


Fig. 3 Please write your figure caption here

The extrinsic parameters relate the rotation R and translation t of the camera origin at the optical center to the world frame. The intrinsic parameters consists of the focal length given by f_x and f_y , the optical center given by c_x and c_y and the skew coefficient s . The distortion parameters consists of radial and tangential distortion along the x and y direction respectively. The number of coefficients are decided based on the lense in consideration. Camera calibration techniques presented in [4] [5] and [6] are widely used in commercial and open source camera calibration tool boxes.

4 What's wrong with monocular camera?

As the name suggests a monocular camera is a camera with a single imaging sensor. In section X. we say that given a perfectly calibrated camera ie. with the intrinsic and extrinsic parameters known, we can measure size of objects in the real world. This however is not true for a single image taken from a monocular camera. When 3D objects are projected onto the image plane (2D), the depth information stored in the z -axis is lost. This can be seen in the example below.

As seen in the above diagram, irrespective of the location of the object in the real world, the size of the object in the image plane remains the same. Mathematically, this can be shown as

$$\frac{y}{f} = \frac{y_1}{z_1 + f} = \frac{y_2}{z_2 + f} \quad (1)$$

Where f is the focal length of the camera. The title of the paper mentions scale estimation. In the context of SLAM which will be covered in the next section, scale is a term used to denote the factor by which the computed trajectory by a SLAM algorithm needs to be multiplied/scaled to make it equal to the ground truth. For example, if the measured distance is 2m, whereas the ground truth is 4m, the scale value would be 2. This is an area of active research and the community has come up with different techniques to solve

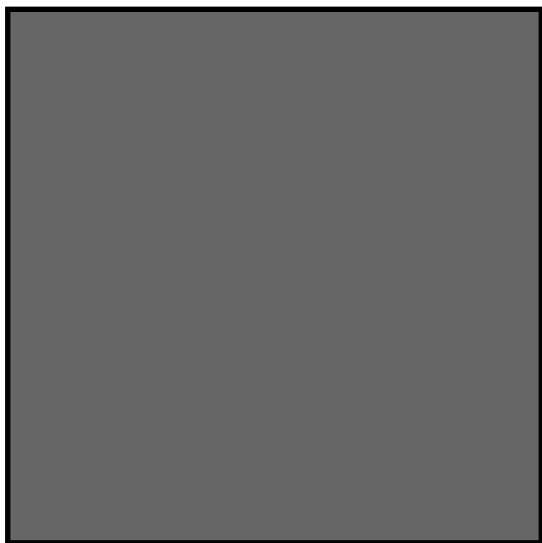


Fig. 4 Monocular camera image coordinates

this problem. As the title suggests, we would be talking about the fusion of IMU and Vision for solving this challenge.

5 Spline Fitting Method

In this paper there are 2 methods of fusing IMU data and Monocular camera. First of which is spline fitting using scale factor with the data of IMU and Monocular camera. Depending on the reliability of the data from both sources to estimate distance. So here we did simulation of whole system with arbitrary value of data considering linear moment of robot with noisy data. These simulations can be seen in result section. With the matlab code. So initially theory behind curve fitting is explained to have a algebraic background of types of curve fitting and their applications. We use spline in curve fitting because we get discrete data from sensors and camera but tracing spline helps for interpolation and extrapolation to have continuous high probability data of path.

Spline is fitted on the data points which is output of sensor in our case. Spline fitting also helps in reducing error due to noise. For Improving accuracy we can increase number of spline but this will also result into increase in computation. So we have to trade off between accuracy and computation. Spline fitting is useful to give smooth curve .N degree spline may has N-1 curves in its shape. So if spline has X3 term that means it can have 2 curves in its spline. Ideal method to cover n points in a plan is by tracing nth order spline which will result into zero error. So spline will pass through all the points and will have a general equation as

$$A_0 + A_1X + A_2X^2 + A_3X^3 + \dots + A_nX^n = Y \quad (2)$$

This problem then converts in simple $Ax = B$ form. we can get values of n terms by plugging in n points values.

Where

$$A = \begin{pmatrix} 1 & X_1 & X_1^2 & \dots & X_1^n \\ 1 & X_2 & X_2^2 & \dots & X_2^n \\ 1 & X_3 & X_3^2 & \dots & X_3^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^n \end{pmatrix} B = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix} C = \begin{pmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad (3)$$

But for practical applications we use other methods and do not plot n degree spline as it is computationally expensive. Secondly if there is more noise in system it will give wrong results. For practical application there are 3 kinds of spline fitting according to the application:-

5.1 Types of curve fitting

5.1.1 Maximum error

This is the method where spline is fit in a way where the point with maximum error is reduced and curve move towards the outlier point. If there exist in order to reduce maximum error of a point. This on the other hand results into

$$E(f) = |\max(f(x_i) - y_i)| \text{ where } i \text{ goes from } 1 \text{ to } n \quad (4)$$

Actually due to one outlier here line shifted in order to reduce maximum error. So this is not reliable process but is easy to compute.

5.1.2 Average error

This method is used to minimize average and fit the spline with minimum error conditions.

$$E(f) = \frac{1}{n} * \sum |f(x_k) - Y_k| \quad (5)$$

Actually due to one outlier here line shifted in order to reduce maximum error. So this is not reliable process but is easy to compute.

5.1.3 RMS (Least square method)

This is the most efficient method to minimize the error this is also called as least square method. Hence this is mostly used in curve fitting. We have also created an algorithm with the help of same type of curve fitting.

$$E(f) = \left\{ \frac{1}{n} * \sum |f(x_k) - Y_k| \right\}^{\frac{1}{2}} \quad (6)$$

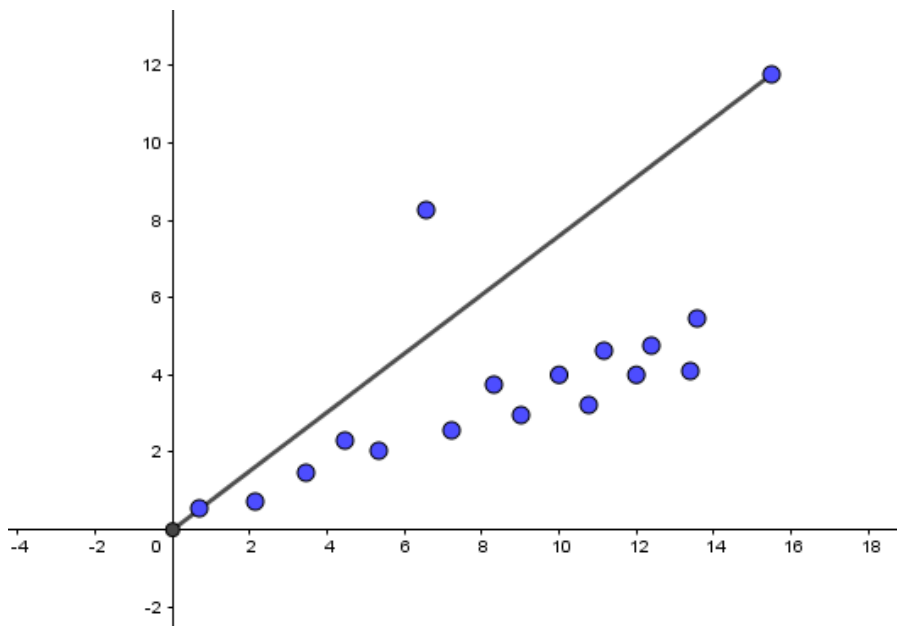


Fig. 5 Curve Fitting with Maximum error

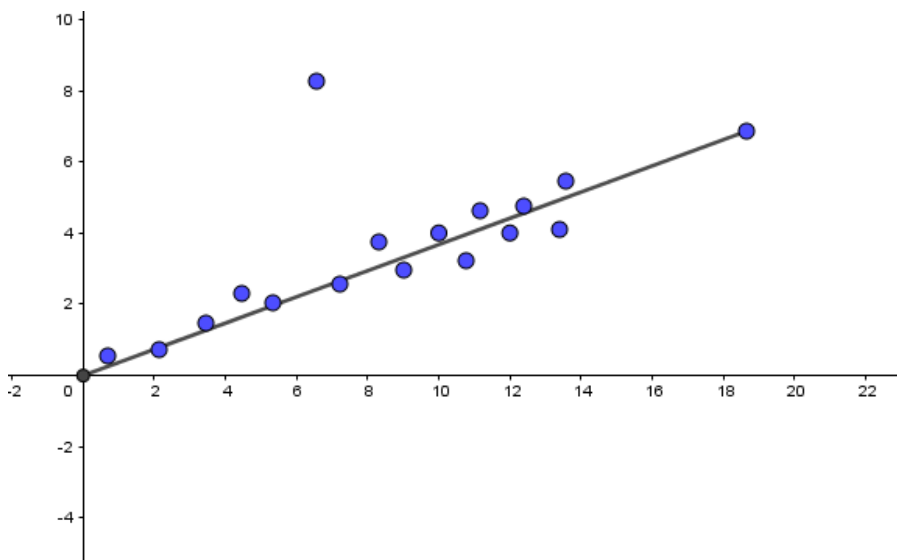


Fig. 6 Curve Fitting with Average error

5.2 IMU Reading

As we don't have real IMU let's simulate virtual readings for this we took X , Y , Z as a function on time

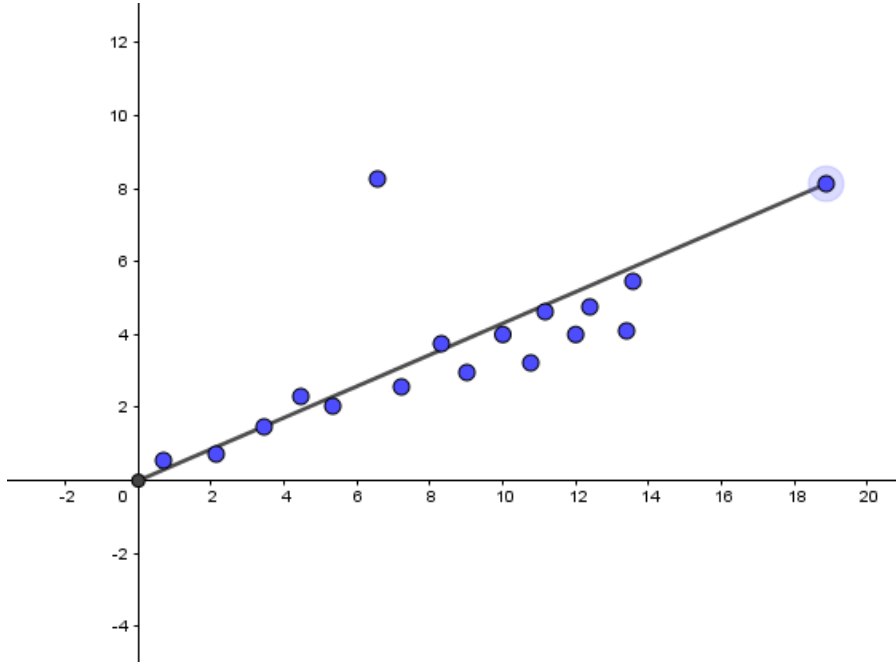


Fig. 7 Curve Fitting with RMS error

$$X_1 = A_y t^2 + B_y t + C_x \quad (7)$$

$$Y_1 = A_y t^2 + B_y t + C_y \quad (8)$$

$$Z_1 = A_z t^2 + B_z t + C_z \quad (9)$$

With this treatment we will add some random noise to all the terms so that the data can be similar to what we obtain from IMU.

$$noise = A * (rand(1, n) - 0.5) \quad (10)$$

So our noise will range from $-A/2$ to $A/2$. As function is $rand()$ so it will change for every point.

$$X = X_1 + noise \quad (11)$$

$$Y = Y_1 + noise \quad (12)$$

$$Z = Z_1 + noise \quad (13)$$

We also took some random points of camera pose. And provided our system with data set. Proof of Least square method in our case: Lets consider optimization of X first.

From equation (7) we have

Here A , B , C are known We have data set of X with respect to t To calculate error in RMS we have

$$Error_x = \left\{ \sum_{i=1}^n (|X_i - (A_x t_i^2 + B_x t_i + C_x)|)^2 \right\}^{\frac{1}{2}} \quad (14)$$

When we differentiate some quantity and equate it to zero it goes to either minima or maxima And if double derivative is positive it goes to minima.

So in this case if we take partial derivatives of error with respect to A_x , B_x , C_x

$$\frac{\partial Error_x}{\partial A_x} = \left\{ \sum_{i=1}^n 2(|X_i - (A_x t_i^2 + B_x t_i + C_x)|) \right\} (-t_i^2) \quad (15)$$

$$\sum_{i=1}^n (|-t_i^2 X_i + (A_x t_i^4 + B_x t_i^3 + C_x t_i^2)|) = 0 \quad (16)$$

$$\frac{\partial Error_x}{\partial B_x} = \left\{ \sum_{i=1}^n 2(|X_i - (A_x t_i^2 + B_x t_i + C_x)|) \right\} (-t_i) \quad (17)$$

$$\sum_{i=1}^n (|-t_i X_i + (A_x t_i^3 + B_x t_i^2 + C_x t_i^1)|) = 0 \quad (18)$$

$$\frac{\partial Error_x}{\partial C_x} = \left\{ \sum_{i=1}^n 2(|X_i - (A_x t_i^2 + B_x t_i + C_x)|) \right\} (-1) \quad (19)$$

$$\sum_{i=1}^n (|-1 X_i + (A_x t_i^2 + B_x t_i + C_x)|) = 0 \quad (20)$$

So we can compute matrix in such a way that the problem changes to

$Ax=B$

Where,

$$A = \begin{pmatrix} \Sigma t^4 & \Sigma t^3 & \Sigma t^2 \\ \Sigma t^3 & \Sigma t^2 & \Sigma t \\ \Sigma t^2 & \Sigma t & \Sigma 1 \end{pmatrix} B = \begin{pmatrix} A_x \\ B_x \\ C_x \end{pmatrix} C = \begin{pmatrix} \Sigma X_i t^2 \\ \Sigma X_i t \\ \Sigma X_i \end{pmatrix} \quad (21)$$

From this equation we can calculate X matrix. Similarly, We can get X matrix of Y and Z axis.

$$A = \begin{pmatrix} \Sigma t^4 & \Sigma t^3 & \Sigma t^2 \\ \Sigma t^3 & \Sigma t^2 & \Sigma t \\ \Sigma t^2 & \Sigma t & \Sigma 1 \end{pmatrix} B = \begin{pmatrix} A_y \\ B_y \\ C_y \end{pmatrix} C = \begin{pmatrix} \Sigma Y_i t^2 \\ \Sigma Y_i t \\ \Sigma Y_i \end{pmatrix} \quad (22)$$

$$A = \begin{pmatrix} \Sigma t^4 & \Sigma t^3 & \Sigma t^2 \\ \Sigma t^3 & \Sigma t^2 & \Sigma t \\ \Sigma t^2 & \Sigma t & \Sigma 1 \end{pmatrix} B = \begin{pmatrix} A_z \\ B_z \\ C_z \end{pmatrix} C = \begin{pmatrix} \Sigma Z_i t^2 \\ \Sigma Z_i t \\ \Sigma Z_i \end{pmatrix} \quad (23)$$

By this treatment we converted given points into equation with least square method here we used condition of

$$\min \begin{pmatrix} A_x t^2 + B_x t + C_x - \lambda_i X_c \\ A_y t^2 + B_y t + C_y - \lambda_i Y_c \\ A_z t^2 + B_z t + C_z - \lambda_i Z_c \end{pmatrix}^2 \quad (24)$$

Here in this formula we have different value of for different spline according to accuracy of camera data and IMU reading.

5.3 Results

In our case we simulated the results and found following outputs. Plotting the second order curve for given data of IMU with just least square method. We get

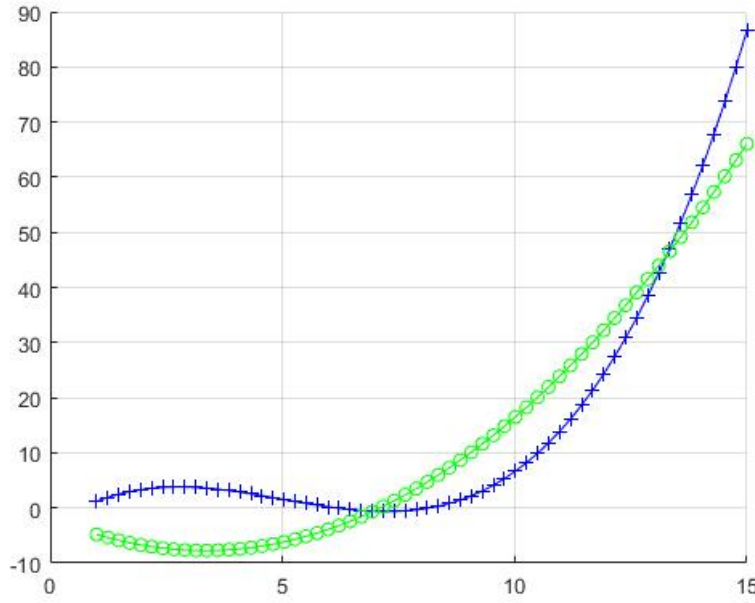


Fig. 8 In this case Blue is actual curve traced by IMU this are points with noise. And green is plotted curve. With least square method.

Then according to the paper we fused data of camera and IMU so we get following graph.

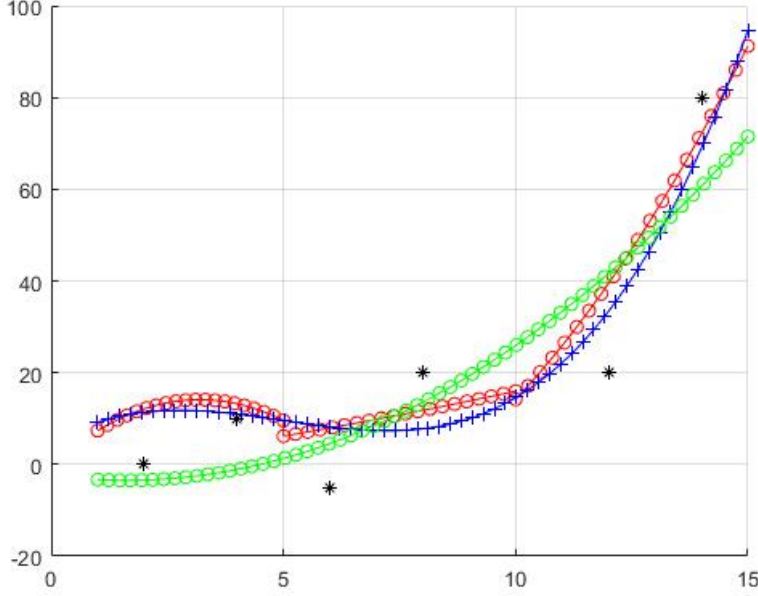


Fig. 9 Multiple spline fit

In Figure 9 we have 3 curved fitted as shown as red which gives better results and it also uses data of camera shown with black dots. So here according to the above formula scale factor of different section is different it is according to the quality of reading by camera. For example, in our case

$$\lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 0.1 \quad (25)$$

In 3rd section camera readings were not accurate so taken into smaller scalar value. This camera data further improved the result. And we get minimum error and good fit. With the given spline fitting method.

So if we look into different of error with traditional Least square method with Jung and Taylor method. It gives good results. Almost we get the same curve.

The graph in Figure 10 has error comparison with and without Jung and Taylor method of split curve fitting and scaling camera input. So it is better to follow the given first method in the paper. It results into considerable reduction in error which is clearly seen in graph.

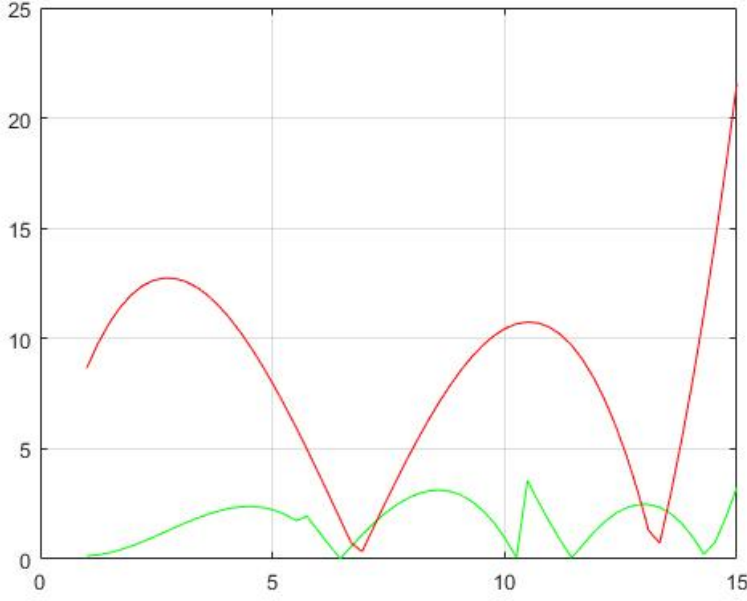


Fig. 10 A. Red curve for method with one curve and least square method. B. Green curve is using the given method by Jung and Taylor.

6 Extended Kalman Filter

6.1 Introduction

In estimation theory, the extended Kalman filter (EKF) is the nonlinear version of the Kalman filter which linearizes about an estimate of the current mean and covariance. In the case of well defined transition models, the EKF has been considered to estimate the state. in this system we have a state and an observer these are given by equation. $X = AX + BU + W_d$ and $Y = CX + W_n$. In this W_d is disturbance and W_n is the noise by sensor. So Kalman filter is used judges between data be disturbance and noise.

If $W_d \gg W_n$ that means disturbance amount is higher than noise hence state estimation can rely more on sensor data for mapping than on previous state. And sensor is more reliable.

if $W_n \gg W_d$ that means noise amount is higher than disturbance hence state estimation can rely more on previous plot data for mapping than on sensor. And state is more reliable than sensor.

So according in paper we learned that non-linear system is defined as.

So now lets take non-linear state equation as follows

$$\vec{z}_{k+1} = \vec{f}_k(\vec{z}_k) + v_k \quad (26)$$

Algorithm 1 Kalman filter algorithm

```

1: Kalman Filter ( $\mu_{t-1}, \Sigma_{t-1}, u_t, z_t$ )
2:  $\mu_t = A_t \mu_{t-1} + B_t u_t$ 
3:  $\Sigma_t = A_t \Sigma_{t-1} A_t^T + R_t$ 
4:  $K_t = \Sigma_t C_t (C_t^T \Sigma_t C_t^T + Q_t)^{-1}$ 
5:  $\mu_t = \mu_t + K_t (z_t - C_t \mu_t)$ 
6:  $\Sigma_t = (I - K_t C_t) \Sigma_t$ 
7: return  $\mu_t, \Sigma_t$ 

```

$$\begin{pmatrix} \vec{x}_{k+1} \\ \vec{v}_{k+1} \\ \vec{a}_{k+1} \\ \vec{\lambda}_{k+1} \end{pmatrix} = \begin{pmatrix} I_3 & \frac{T}{\lambda} I_3 & \frac{T^2}{2\lambda} I_3 & 0 \\ 0 & I_3 & T I_3 & 0 \\ 0 & 0 & I_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{x}_k \\ \vec{v}_k \\ \vec{a}_k \\ \vec{\lambda}_k \end{pmatrix} \quad (27)$$

where $x_k + 1$ is the position without scale of the IMU/Camera and v_{k+1} , $a_k + 1$ are the velocity and acceleration of the IMU/Camera in metric unit [m]. \vec{z}_k is the gaussian process noise. Every vector in \vec{z}_k is resolved in the world frame W. Note that we do not include the orientation information in the model nor use it as a measurement in order. to keep the algorithm simple and fast. On each acceleration measurement we do the conversion from the inertial to the world frame by using a zero order hold of the unfiltered attitude measurement returned by the visual SLAM framework. As we work in a middle size environment with enough loops we assume negligible drift in the SLAM map and assume thus highly accurate attitude estimation from the visual SLAM framework. The model in its linearized form yields,

$$F_k = \begin{pmatrix} I_3 & \frac{T}{\lambda} I_3 & \frac{T^2}{2\lambda} I_3 & -\frac{T}{\lambda^2} I_3 & \frac{T^2}{2\lambda^2} I_3 \\ 0 & I_3 & T I_3 & 0 & 0 \\ 0 & 0 & I_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (28)$$

For fusion implementation we consider the measurements in different observation vectors. For a multi rate filter, as it is in our case, the literature suggests two solutions. One would be using a (higher order) hold to synchronize the different measurements. Another is to weight the uncertainty of the measurement according to its temporal occurrence. We claim no certainty at all if no measurement is available (i.e. the measurement noise variance is infinite). Thus the update equations simplify to improve results. A more complex weighting function (i.e. exponential decay in time) could also be applied, however, at the cost of speed. The measurement updates for the vision and the IMU yields (V and I denotes Vision and IMU)

The innovation for the vision part is,

$$K_{V,k} = P_k^- H_{V,k}^T (H_{V,k} P_k H_{V,k}^T + R_v)^{-1} \quad (29)$$

$$\vec{z}_k = \vec{z}_k + K_{V,k} (\vec{x}_{SLAM} - H_{V,k} \vec{z}_k) \quad (30)$$

$$P_k = (I - K_{V,k} H_{V,k}) P_k^- \quad (31)$$

The innovation for the IMU part is,

$$K_{I,k} = P_k^- H_{I,k}^T (H_{I,k} P_k^- H_{I,k}^T + R_I)^{-1} \quad (32)$$

$$\vec{z}_k = \vec{z}_k + K_{I,k} (\vec{x}_{IMU} - H_{I,k} \vec{z}_k) \quad (33)$$

$$P_k = (I - K_{I,k} H_{I,k}) P_k^- \quad (34)$$

The two matrices R_I , R_V are the noise covariance matrices for the vision and IMU measurement inputs x_{SLAM} , a_{IMU} which are resolved in the world frame W . The vector x_{SLAM} is the position without scale obtained from the vision algorithm (SLAM). The IMU measurement a_{IMU} needs special attention, because significant errors arise in the conversion from the raw IMU output.

$$\vec{a}_w = R_{wc} R_{ca} (\vec{a}_a - \vec{b}) - \vec{g}_w \quad (35)$$

7 Simultaneous Localization and Mapping

In this section, we cover SLAM, and the different concepts related to SLAM such as Visual Odometry. We then explore the two different approaches to SLAM mentioned in the paper. We briefly cover filter based SLAM called Extended Kalman Filter SLAM (EKFSLAM) and a key-frame based SLAM called Parallel Tracking and Mapping (PTAM) [6]. Simultaneous Localization and Mapping (SLAM) is a technique for estimating the motion of the robot and reconstructing the map/structure of the unknown environment. SLAM using only visual information only is specifically referred to as visual SLAM (vSLAM). The SLAM problem can be stated as follows:

How can a body navigate in a previously unknown environment while constantly building and updating a map of its workspace using onboard sensors only? [9]

Mathematically, this can be expressed as [12]

Input:

$$\text{Control input : } u_{1:T} = u_1, u_2, \dots, u_T \quad (36)$$

$$\text{Observations : } z_{1:T} = z_1, z_2, \dots, z_T \quad (37)$$

Output:

$$\text{Map of Environment : } m \quad (38)$$

$$\text{Path of robot : } x_{0:T} = x_0, x_1, \dots, x_T \quad (39)$$

From the problem statement above, we notice that the robot has no a priori knowledge of the workspace or environment that it is in. This makes SLAM a very challenging problem in probabilistic robotics. This is also referred to as a chicken-egg problem in that we need to map the environment to get an accurate pose, but at the same time we also need an accurate pose to build a correct map. Therefore, it is an iterative process of estimating pose and

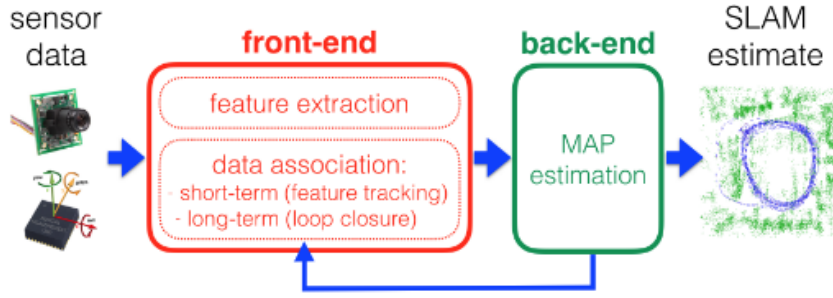


Fig. 11 SLAM architecture

building a map simultaneously. Accurate pose estimation is critical for many applications in computer vision, autonomous robotics and augmented reality.

From Fig. 11 we see that a typical SLAM system consists of a front-end and a back-end. The front-end is responsible for feature extraction while the back-end is responsible for MAP estimation. Both the methods (MonoSLAM & PTAM) are feature based implementations.

7.1 Visual Odometry

Odometry is the technique of estimating the position of a robot over time using sensors such as cameras, wheel encoders or any sensor measuring relative movement. Compared to SLAM which maintains a global consistent map, visual odometry (VO) maintains a local consistent map optimized over the last n frames. A generalized VO pipeline can be summarized as follows

Different algorithms exist due to the different type of cameras available such as Stereo, Mono, RGB-D. As we use a monocular setup in our research paper, we would be going through the visual odometry algorithm from 2D to 2D correspondences for a monocular sensor. The algorithm is summarized as follows.

Algorithm 2 Monocular Visual Odometry

- 1: Capture a new image frame I_k
 - 2: Extract and match features between frames I_k and I_{k-1}
 - 3: Compute the essential matrix E from the matched feature points
 - 4: Decompose E into R_k and t_k to form T_k
 - 5: Choose the correct T_k matrix and scale t_k accordingly from scale factor
 - 6: Compute $C_k = C_{k-1} * T_k$
 - 7: Goto step 1.
-

Since we are using a monocular sensor, we always compare the image I_k captured at time instant k with the image I_{k-1} capture at time instant $k - 1$. In a stereo setup, we would compare images from the left and right part of the

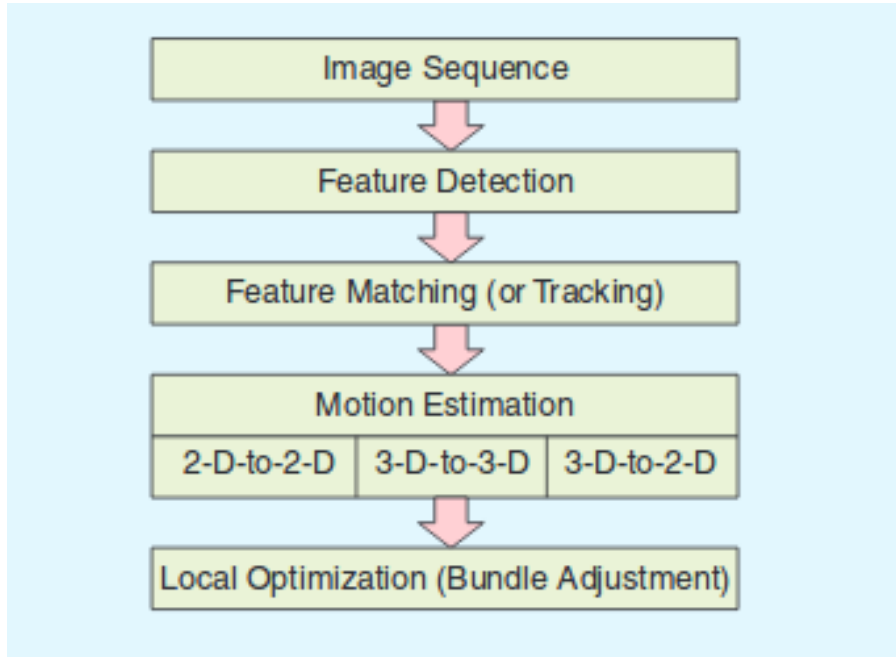


Fig. 12 Generalized Visual Odometry Pipeline

stereo. Once an image is captured, we compute the features for image I_k and I_{k-1} . In the feature detection step, the image is searched for key-points called features which are likely to be present in successive images. Features can be defined as image patterns that differ from its immediate neighborhood in terms of intensity, color and texture. There exist a variety of feature detectors in literature which vary in performance and properties. Some of the properties of feature detectors are rotation, scale and affine invariant, repeatability, localization accuracy, robustness and efficiency. Heitanen et al. present a comparison of feature detectors and descriptors for object class matching in [3]. Once we detect features, we then need to compute a descriptor such that features detected in two images can be compared with each other. The easiest approach would be to form a patch of pixels and compare using a sum of squared distances (SSD) or normalized cross correlation (NCC) metric. Such descriptors are not invariant to any of the above mentioned properties and perform poorly in practical applications. One of the popular descriptors called SIFT [8] uses gradient orientations as its descriptors. This forms a 128-element descriptor that is invariant to most of the above mentioned properties such as rotation, scale and illumination which makes it widely applicable in practical real-time applications. For extracting features from two images there are two paths with one being to detect and match features independently in two frames and the other being to track features in subsequent frames an example of which is a KLT Tracker [13]. There are various methods employed to match features

accurately such a RANSAC which stands for Random Sample Consensus and is used to remove outliers among feature matches. Once we detect and match features in I_k and I_{k-1} , we calculate the essential matrix E which describes the geometric relationship between two images. We can use Nisters five point algorithm [10] or Longuet-Higgins eight point algorithm [7] to get the essential matrix E . Once we get the essential matrix E , we decompose it to extract the rotation and translation parts. Four different solutions are obtained of which the correct pair can be found out by triangulation. The solutions are given as

$$R = U(\pm W^T)V^T \quad (40)$$

$$t = U(\pm W^T)SV^T \quad (41)$$

where R is the rotation matrix and t is the translation vector.

7.2 MonoSLAM

MonoSLAM was developed by Davison et al. [1, 2]. It uses EKF as an estimator. In EKFSLAM, the state \mathbf{x} space is represented by the robot state \hat{x}_ν along with the position of the different landmarks \hat{y}_i that the robot observes. The state vector along with the covariance matrix can be represented as follows

$$\hat{x} = \begin{pmatrix} \hat{x}_\nu \\ \hat{y}_1 \\ \hat{y}_2 \\ \vdots \end{pmatrix} P = \begin{pmatrix} P_{xx} & P_{xy1} & P_{xy2} & \cdots \\ P_{y1x} & P_{y1y1} & P_{y1y2} & \cdots \\ P_{y2x} & P_{y2y1} & P_{y2y2} & \cdots \\ \vdots & \vdots & \vdots & \cdots \end{pmatrix} \quad (42)$$

The covariance matrix is used to represent the first order uncertainty. As the robot moves around, new features are added to the state vector. It is important to note that the robot initialization is done by using a known object. The prediction and correction steps are performed as mentioned in the EKF algorithm in section 6. One of the problems with this method is that the state vector becomes large when the number of land marks or key-points tracked increases. This increases the computation time.

7.3 Parallel Tracking and Mapping (PTAM)

PTAM is a SLAM system specifically designed to track a hand-held camera in a small AR workspace. For real-time operation, PTAM splits the tracking and mapping threads into two separate tasks running in parallel. The tracking thread robustly tracks hand-held motion and the mapping thread produces a 3D map of the features. Although developed specific to AR applications, the concept of running operations on multiple threads was incorporated into future SLAM algorithms.

The map consists of a collection of key-points/features that are located in a world frame W . They are represented in the homogeneous form. The map also contains key-frames which are instances of the frames taken from the hand-held camera. Each point is stored with a source key-frame. Unlike earlier methods which process each and every frame, this method only processes frames when there is sufficient information. Thus incremental mapping is replaced with a more computationally expensive batch method called bundle adjustment.

The tracking algorithm can be summarized as follows.

Algorithm 3 PTAM - Tracking Algorithm

- 1: A new frame is captured and a prior pose estimate is generated from motion model.
 - 2: Map points are projected into the image based on prior pose.
 - 3: Small number of features (50) are searched in the image.
 - 4: The camera pose is updated from the feature matches.
 - 5: Large number of points is re-projected and searched in the image.
 - 6: Final pose estimate is computed from all matched found.
 - 7: Goto step 1.
-

PTAM is initialized by a lateral offset movement of the hand-held camera. This can be assumed as a stereo image for starting the mapping. The initial map is constructed using the 5 point algorithm [10] with arbitrary scale. As the camera moves, the key-frames increase from an initial two. Here, we only add key-frames if the tracking quality is good and a minimum of twenty frames has passed from the last keyframe. Bundle adjustment is iteratively performed to adjust the map based on a cost function. The mapping algorithm can then be summarized as.

Algorithm 4 PTAM - Mapping Algorithm

Require: Stereo Initialization

- 1: **if** New Keyframe **then**
 - 2: Update key-frame
 - 3: Integrate key-frame
 - 4: Add new features
 - 5: **else if** Locally Converged **then**
 - 6: **if** Globally Converged **then**
 - 7: Update data association
 - 8: **else**
 - 9: Global bundle adjust
 - 10: **end if**
 - 11: **else**
 - 12: Locally bundle adjust
 - 13: **end if**
 - 14: Sleep 5ms
 - 15: Goto step 1.
-

Finally when compared with Filter based SLAM, PTAM can handle thousands of features by splitting the tracking and mapping onto different threads

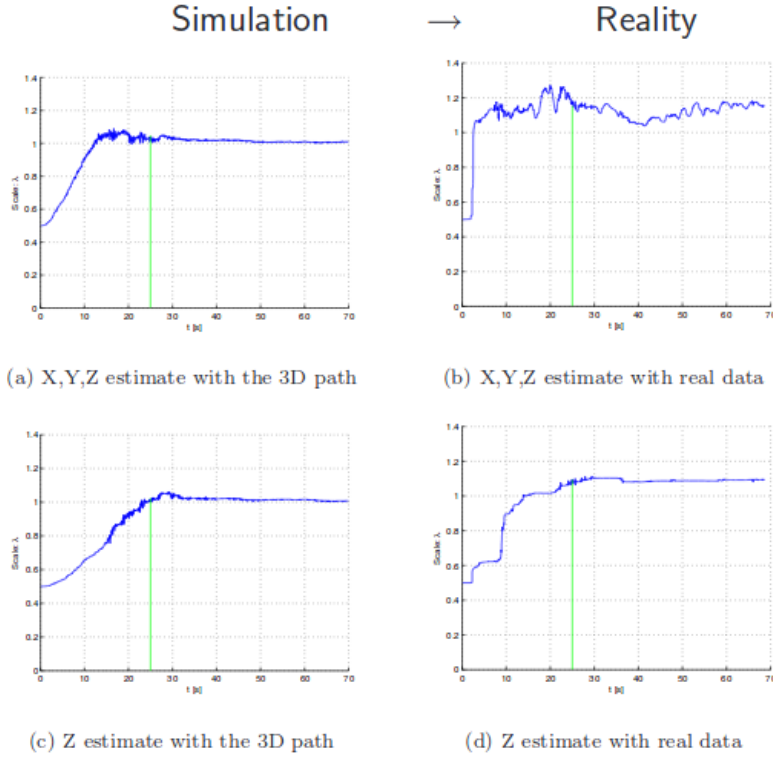


Fig. 13 Offline simulation results with simulated and real data. Fewer directions (X,Y,Z) included in Kalman filter gives us a better estimate of λ . [11]

on the CPU. We see that SLAM algorithms built now-a-days have used this philosophy to great success.

7.4 Results

7.4.1 Simulated and Real Data

In the paper, the Kalman filter was simulated offline with simulated data from the spline fitting section. The two inputs to the Kalman filter were the position from the vision sensor and acceleration from the IMU. Three different approaches were taken. The first method is same as Eq. 7, the second method only uses the Z-axis which gives the states as $[\vec{x}_z, \vec{v}_z, \vec{a}_z, \lambda]$ and the third uses only the X and Y-axis.

Figure 13 shows the scale estimation $\lambda(t)$ for a simulated 3D path on left and actual path on right. The standard deviation for the acceleration noise for simulation was chosen same as real data ($\sigma_{SLAM} = 0.01, \sigma_{IMU} = 0.2m/s^2$) with initial velocity and acceleration set to zero. Plots in Fig. 13a and Fig. 13c

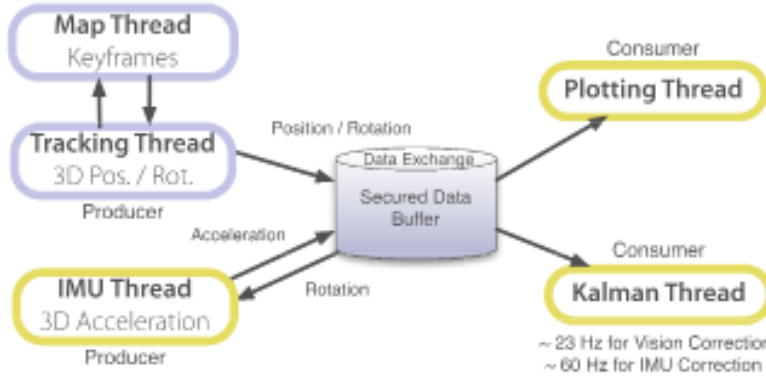


Fig. 14 Online implementation with PTAM work flow. Blue color boxes are original PTAM threads. Yellow boxes are added threads for scale estimation λ . [11]

do not differ due to simulating the acceleration from the spline ideally in world frame. The estimate becomes more accurate when we use less number of directions (X,Y,Z). Wrong measurements in acceleration influence λ and make the scale estimation sensitive as seen in Fig. 13b. Hence only using a single axis (Z-axis) gives us the best result.

7.4.2 Online Implementation

For the online implementation, the third setup was incorporated into the PTAM code [5]. As mentioned in section 7.3 PTAM employs two parallel threads called Tracker for tracking and MapMaker for mapping. Two additional threads for IMU and Kalman we added. The IMU thread provides the acceleration measurements. The Kalman thread starts with λ calculated from integrating acceleration values, position from SLAM algorithm and acceleration and velocity set to 0. Values for covariance matrix \mathbf{Q} is time-varied which provides control over the sensitivity of the Kalman filter. For our report we have not incorporated the online implementation and have reported results from the paper.

References

1. Andrew J Davison. Real-time simultaneous localisation and mapping with a single camera. page 1403. IEEE, 2003.
2. Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1052–1067, 2007.
3. Antti Hietanen, Jukka Lankinen, Joni-Kristian Kämäräinen, Anders Glent Buch, and Norbert Krüger. A comparison of feature detectors and descriptors for object class matching. *Neurocomputing*, 184:3–12, 2016.
4. S-H Jung and Camillo J Taylor. Camera trajectory estimation using inertial sensor measurements and structure from motion results. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2001.
5. Georg Klein and David Murray. Source code of ptam (parallel tracking and mapping). URL: <http://www.robots.ox.ac.uk/~gk/PTAM/>. Last visited on 11/18/2018.
6. Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.
7. H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133, 1981.
8. David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
9. Roland Siegwart Margarita Chli, Martin Rufli. Lecture notes in autonomous mobile robots, 2017. URL: https://www.ethz.ch/content/dam/ethz/special-interest/mavt/robotics-n-intelligent-systems/asl-dam/documents/lectures/autonomous_mobile_robots/spring-2017/2017%20-%20SLAM%20I%20-%20after%20lecture.pdf. Last visited on 11/18/2018.
10. David Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004.
11. Gabriel Nützi, Stephan Weiss, Davide Scaramuzza, and Roland Siegwart. Fusion of imu and vision for absolute scale estimation in monocular slam. *Journal of intelligent & robotic systems*, 61(1-4):287–299, 2011.
12. Cyrill Stachniss. Lecture notes in robot mapping, 2013. URL: <http://ais.informatik.uni-freiburg.de/teaching/ws12/mapping/pdf/slam04-ekf-slam.pdf>. Last visited on 11/18/2018.
13. Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. 1991.
14. Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22, 2000.