



## Motion and Saliency Based Monocular SLAM

AHMINE Yassine\* , CHOUIREB Fatima

*Electronic department, Laghouat University, Telecommunications, Signals and Systems Laboratory.*

### **Article history**

Submitted date: 2017-07-05

Acceptance date: 2017-12-03

### **Abstract**

Feature extraction is a key component of a Monocular Simultaneous Localization and Mapping (Monocular SLAM) system, which permits to extract features that can be reliably tracked over frames. This paper proposes a novel approach for Monocular SLAM that uses the information on the camera displacement and image saliency to adequately extract stable features, which will be prompt to produce sufficient parallax that is essential to ensure precise localization and mapping. The results obtained from real data show that the proposed method outclasses the state of the art method both in precision and computational speed.

**Key-words :** *Monocular SLAM ; EKF-SLAM ; FAST Detector ; Visual Saliency.*

### **Résumé**

L'extraction des amers est un élément clé d'un système de localisation et de cartographie monoculaire simultanée (Monocular SLAM), qui permet d'extraire des points d'intérêts qui peuvent être suivis de manière fiable au travers des images. Cet article propose une approche originale pour le SLAM Monoculaire qui utilise l'information sur le déplacement de la caméra et la saillance de l'image pour extraire adéquatement des points d'intérêts stables, qui seront plus prompts à produire suffisamment de parallaxe, élément essentiel pour assurer précisément la localisation et la cartographie. Les résultats obtenus à partir de données réelles montrent que la méthode proposée surpasse la méthode de l'état de l'art en termes de précision et de vitesse de calcul.

**Mots-clés :** *SLAM Monoculaire ; EKF-SLAM ; Détecteur FAST ; Saillance Visuelle*

\* Corresponding author. Tel. : +213 555520174.

E-mail address: [yacineahmine@gmail.com](mailto:yacineahmine@gmail.com).

## 1. Introduction

Simultaneous Localization and Mapping (SLAM) is one of the key problems in robotics. It is a typical chicken and egg problem; where to achieve accurate localization, a precise map is necessary, and to achieve precise mapping, an accurate knowledge of the robot locations is needed. To solve this problem, an Extended Kalman filter can be used to jointly estimate the robot and landmarks (map) locations by exploiting the information provided by the measurements at different locations.

Researches involving using a single camera as the only sensor to perform SLAM, known as Monocular SLAM, gained a lot of attention over the past fifteen years because of the advantages of the cameras over other sensors, like laser range finders, in terms of price and power consumption.

In his paper, Davison [1] presented the first implementation of a SLAM system that used a camera as the only sensor. This work opened the way to a number of works on Monocular SLAM. Notably, the work proposed by Eade and Drummond [2], which used a graph of local frames to avoid map inconsistency and the work of Civera et al. [3] that integrated both the RANSAC algorithm for outlier's rejection and the inverse depth parameterization [4]. Other approaches that consider the Kalman filter framework can be found on literature like the system proposed by Lee [5], which combined a particle filter and an unscented Kalman filter.

In the context of camera based SLAM, feature extraction plays a key role in the ability of the system to extract robust features and track them during the exploration of the environment. Frintrop et al. [6] proposed a strategy inspired by the human visual intention system [7] to extract a sparse set of features. This strategy was later used in combination with learned objects database by Kuan-Ting Yu et al. [8] to select key features.

In this paper, we propose a method where features are not uniformly distributed in the image like in the work of Civera [3], but where they are distributed according to the camera displacement and the salient

regions in the image, in order to extract more robust features and augment the performances of the system.

This paper is organized as follows: first, the proposed method is described, and then the EKF Monocular SLAM based on inverse depth is reviewed. After that, the results are presented and a comparison between the proposed method and the state-of-the-art monocular SLAM algorithm of Civera et al. [3], is done. The conclusions are given in the last section.

## 2. Method description

The proposed method aims to determine, according to camera displacement and image saliency zones, parts of the image that are more suitable to initialize new features. This is done by dividing the input image into 8 zones as shown in Fig. 1. The reason why using 8 zones is to compromise between the accuracy "induced by the augmentation of their number" and the computational cost inherent to this augmentation.



Fig. 1. Image division into 8 zones

To each zone ( $z_i ; i = 1, \dots, 8$ ) is attributed a coefficient  $c_i$  which is computed according to two other coefficients  $m_i$  and  $s_i$ . The coefficients  $m_i$  are calculated based on camera motion to provide a motion based grid. The coefficients  $s_i$  are calculated based on saliency of the corresponding zone to provide a saliency based grid.


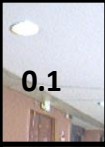

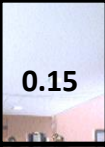




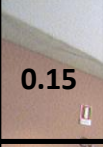
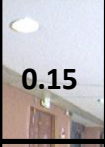












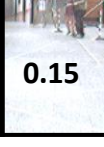

				Straight
0.15	0.1	0.1	0.15	
				Straight
0.15	0.1	0.1	0.15	
				Turning Left
0.15	0.15	0.1	0.1	
				Turning Left
0.15	0.15	0.1	0.1	
				Turning Right
0.1	0.1	0.15	0.15	
				Turning Right
0.1	0.1	0.15	0.15	

Fig. 2. Grid coefficients applied to every type of camera displacement

#### a. Motion Based Grid

This approach permits to take advantage of the information on the type of motion. To give more importance to the zones that are more likely to provide useful features. Since the benchmarking of the proposed method was done using the RAWSEEDS dataset [9], only three types of displacement were considered: straight, turning left and turning right. To each type of these displacements is associated a specific grid as presented in Fig. 2.

These coefficients were attributed based on the following considerations: in the case of a straight displacement, features that are located on image ends are more suitable than features located on image center because they will produce more parallax that is essential for good performance. For a camera turning left, features located on the right half of the image will be rapidly unusable by the SLAM system. That's makes

them undesirable for initialization. The right turning case can be conversely considered.

#### b. Saliency Based Grid

In this approach a saliency map is used to compute the coefficients  $s_i$ . The saliency map is generated using VOCUS [10], a method inspired by human intention system that uses variations on intensity, orientation and color to determine regions of interest (salient regions) in the image. The saliency map produced is a gray level image, where the brightest regions represent the most salient ones. The  $s_i$  coefficients are the mean value of the corresponding zone  $z_i$  on the saliency map.

$$s_i = \frac{1}{H.W} \sum_{i=i_0}^H \sum_{j=j_0}^W S(i,j) \quad (1)$$

Here H represents the zone height and W represents the zone width.  $S(i,j)$  is the value of the saliency map on the location specified by  $i$  and  $j$ .  $(i_0, j_0)$  are the location of the up left zone corner. Fig. 3 shows an example of a saliency map.



Fig. 3. Saliency map of the input image

#### c. Fusing Grids

To compute the  $c_i$  coefficients the two maps are fused by multiplying each corresponding  $m_i$  and  $s_i$  coefficients.

$$c_i = \eta \cdot m_i \cdot s_i \quad / \quad i = 1, \dots, 8 \quad (2)$$

Where  $\eta$  is a normalizer which ensures that the sum of the coefficients  $c_i$  is equal to one. This result is used to determine how many features, from the overall number of features that have to be initialized, should be extracted from each zone.

$$n_i = \text{round}(N \cdot c_i) \quad (3)$$

Where  $n_i$  is the number of extracted features in  $z_i$  and  $N$  the total number of extracted features.

### 3. Monocular ekf-slam based on inverse depth

In this section the inverse depth based monocular SLAM [4] (Mono-SLAM) is reviewed.

#### a. Camera Motion Model

The state vector is defined as follows:

$$X_{k|k} = [x_{k|k}^c, L_{k|k}]^T \quad (4)$$

Where the camera state vector is  $x_{k|k}^c$ :

$$x_{k|k}^c = [r^{cw}, q^{cw}, v^{cw}, w^{cw}]^T \quad (5)$$

$r^{cw}$  is the camera optical center position and  $q^{cw}$  represents the camera orientation in the world reference frame.  $v^{cw}$  and  $w^{cw}$  are the camera linear and angular velocities respectively. The camera motion model  $f_v()$  is:

$$x_{k+1|k}^c = f_v(x_{k|k}^c) \quad (6)$$

And

$$x_{k+1|k}^c = \begin{bmatrix} r^{cw} + (v^{cw} + V^{cw}) \cdot \Delta t \\ q^{cw} \times q((w^{cw} + W^{cw}) \cdot \Delta t) \\ v^{cw} + V^{cw} \\ w^{cw} + W^{cw} \end{bmatrix} \quad (7)$$

Where  $V^{cw}$  and  $W^{cw}$  are zero mean Gaussian distributions representing constant linear and angular velocities

#### b. Inverse Depth Parameterization

The feature state  $l_i$  is a 6D vector:

$$l_i = [x_i \ y_i \ z_i \ \rho_i \ \theta_i \ \varphi_i]^T \quad (8)$$

where  $[x_i \ y_i \ z_i]$  represents the camera optical center position from which the interest point was first detected,  $\rho_i = \frac{1}{d_i}$  represents the inverse depth of the interest point, and  $\theta_i, \varphi_i$  represent its azimuth and elevation (encoded in the world reference frame) which permits to define the directional vector  $m(\theta_i, \varphi_i)$ :

$$m(\theta_i, \varphi_i) = [\cos(\varphi_i) \sin(\theta_i) \ -\sin(\varphi_i) \ \cos(\varphi_i) \cos(\theta_i)]^T \quad (9)$$

#### c. Measurement Model

In this paper the pinhole model is used. Thereby, the image coordinates of a projected feature  $l_i$  on the image plane are:

$$y_i^h = \begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} u_0 - f \frac{h_x^c}{h_z^c} \\ v_0 - f \frac{h_y^c}{h_z^c} \end{bmatrix} \quad (10)$$

Where  $u_0, v_0$  represent the camera's principal points.  $f$  is the focal length expressed in pixel units.  $h_x^c, h_y^c, h_z^c$  are the coordinates of the feature  $l_i$  expressed in camera frame. The transformation from the world reference frame to the camera frame is expressed by the following relation:

$$\begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} = R^{cw} \left( \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + \frac{1}{\rho_i} m(\theta_i, \varphi_i) - r^{cw} \right) \quad (11)$$

Here  $R^{cw}$  represents the rotation from the world reference frame to the camera frame.

The camera distortion model used is:

$$\begin{bmatrix} u_d \\ v_d \end{bmatrix} = \begin{bmatrix} \frac{u - u_0}{1 + K_1 r^2 + K_2 r^4} + u_0 \\ \frac{v - v_0}{1 + K_1 r^2 + K_2 r^4} + v_0 \end{bmatrix} \quad (12)$$

Where  $K_1$  and  $K_2$  are the distortion coefficients.

#### d. Kalman Filter Framework

The motion model used for the prediction step of the Kalman filter is:

$$X_{k+1|k} = \begin{bmatrix} f_v(x_{k|k}^c) \\ L_i \end{bmatrix} \quad (13)$$

With the predicted covariance matrix:

$$P_{k+1|k} = F P_{k|k} F^T + G P_{k|k} G^T \quad (14)$$

Where  $F$  and  $G$  represent the Jacobian matrices of the motion model with respect to (w.r.t.) the state vector  $X$  and the Gaussian distribution  $n = \begin{bmatrix} V^{cw} \\ W^{cw} \end{bmatrix}$  respectively.

The update of the filter is done according to the following equations:

$$x_{k+1|k+1} = x_{k+1|k} + K(Z^h - Y^h) \quad (15)$$

And

$$P_{k+1|k+1} = P_{k+1|k} - KSK^T \quad (16)$$

With

$$S = HP_{k+1|k}H^T + R \quad (17)$$

$$K = P_{k+1|k}H^T S^{-1} \quad (18)$$

Where  $H$  is the Jacobian matrix of the measurement model w.r.t. the state vector.  $R$  is the covariance matrix of the measurement process noise.

#### 4. Results

The proposed method was tested on MATLAB, using the RAWSEEDS dataset (Bicocca-2009-02-25b), that provides multiple sensor streams from which only the monocular sequence was used for the SLAM. This sequence was recorded from a traveling robot around the Università di Milano-Bicocca, in Milan (Italy). What makes this dataset challenging is that it contains low textured parts (corridor) and dynamic elements (people). The detector used in the experiments is the FAST detector [11]. The results presented in this section were obtained using the benchmarking solution provided in free download on the RAWSEEDS web page.

##### a. The state-of-the art Method

Fig. 4 shows the trajectory estimated using the approach of Civera et al [3] for the monocular SLAM, and the ground truth.

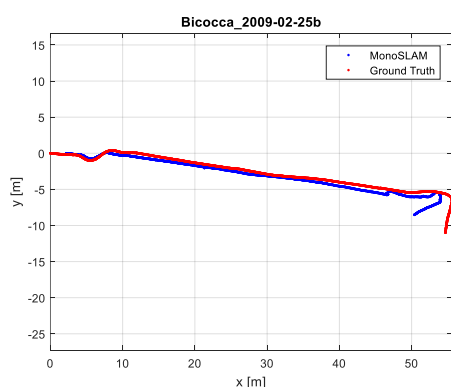


Fig. 4. Estimated trajectory using the state of the art method (blue) and ground truth (red)

The absolute trajectory error (ATE) is presented in Fig. 5.

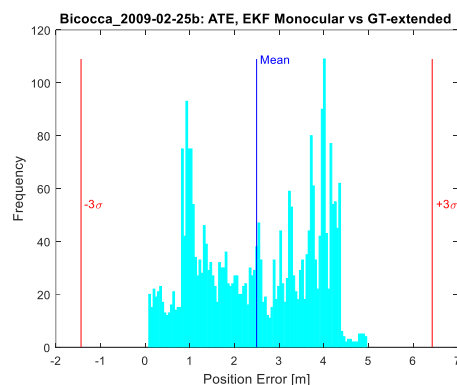


Fig. 5. The absolute trajectory error (the mean is showed in the center of the figure)

From these figures, we can see that the state-of-the-art method provides a relatively precise estimation of the robot trajectory, while consuming an important time for the execution (9206.247s overall).

##### b. The Saliency Based Method

In this experiment, only the saliency information was used for the monocular SLAM, and the system failed to continue the estimation beyond 10 meters. The trajectory estimation is presented with the ground truth on the following figure.

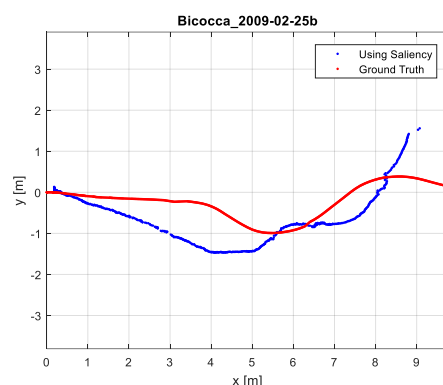


Fig. 6. Estimated trajectory using only the saliency information method (blue) and ground truth (red)

The execution time for every iteration is showed in Fig.8.

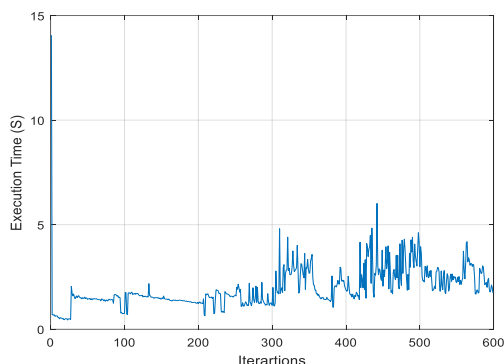


Fig. 7. Evolution of the execution time during the experiment

Here, the results show a bad estimation of the trajectory, because during navigation, features located only on salient regions can exhibit insufficient parallax which is not suitable for good localization and mapping.

*c. The Proposed Method*

The method proposed in this work uses the information of both the trajectory and the saliency and exhibits the following results.

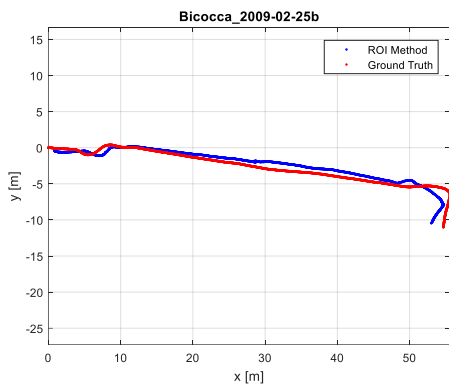


Fig.8. Estimated trajectory using the proposed method (blue) and ground truth (red)

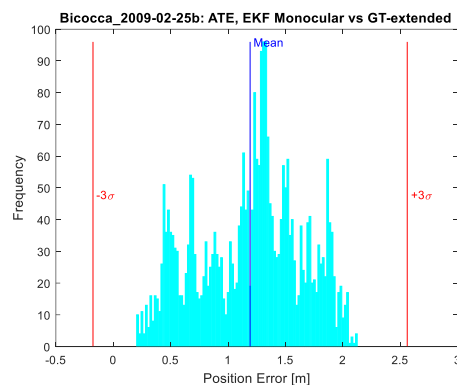


Fig. 9. Occurrence frequency of the different values of the ATE (the mean is showed in the center of the figure)

Fig. 10 shows the execution time for both the state-of-the-art and the proposed methods.

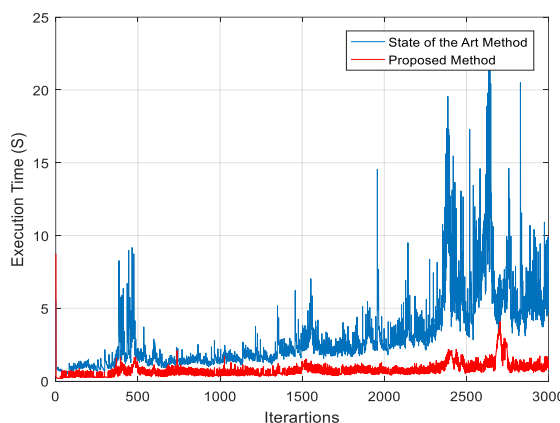


Fig.10. Evolution of the execution time for both the state of the art (in blue) and the proposed methods (in red)

From these figures we can see that the strategy based on using both the trajectory and saliency information gives the best performances in terms of precision and execution time. Taking advantage from this information, the proposed strategy permits to use less features for the SLAM (50 features on average) than the state of the art method (110 features on average), and thereby induces a decrease in execution time. Fig.11 shows the evolution of the number of features used by the two methods. It also permits to exploit the features in a more useful way because they will be located in zones (regions) that are more suitable for the localization and mapping, and inducing by this mean a gain in precision in comparison with the state-of-the-art and saliency based methods.

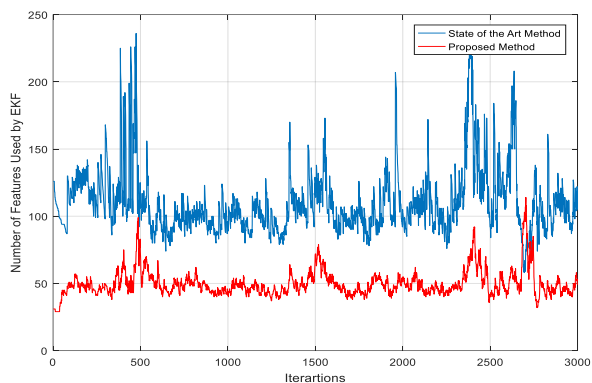


Fig.11. Evolution of the number of features handled by the EKF for both methods

## 5. Conclusion

In this work, a practical method based on the extended Kalman filter (EKF) was presented and tested using real data from a navigating robot on a dynamic environment. The proposed strategy uses the camera displacement and the image saliency information to determine the number of features to be initialized in every image zone. Which induces a reduction of the number of features handled by the EKF (by approximately 55%) and by this mean a reduction of the computational time, and permits to reach a smaller absolute error on the trajectory than the state of the art method. This permits to say that the method presented in this paper outclasses the state of the art method in terms of precision and execution speed.

Improvements for the presented method can be studied. Notably the tracking of the salient regions over the frames to reduce the computational cost inherent to their generation. Other detectors can be considered (like the SURF detector) to test their impact on the SLAM performances.

## 6. Acknowledgment

Authors would like to thank J. Civera for providing the code of the “1-Point RANSAC Inverse Depth EKF Monocular SLAM”.

## References

- [1] J. Davison. Real-time simultaneous localisation and mapping with a single camera. Proceedings of the 9th International Conference on Computer Vision 2003, Nice.
- [2] E. Eade, T. Drummond. Monocular SLAM as a graph of coalesced observations. Proceedings of the International Conference on Computer Vision (ICCV) 2007.
- [3] Javier Civera, Óscar G. Grasa, Andrew J. Davison, J. M. M. Montiel. 1-Point RANSAC for EKF Filtering: Application to Real-Time Structure from Motion and Visual Odometry. Journal of Field Robotics 2010; 27(5):609-631.
- [4] J. Montiel, J. Civera, and A. J. Davison. Unified inverse depth parametrization for monocular SLAM. Robotics Science and Systems Conference 2006, Philadelphia.
- [5] Lee S. Real-time camera tracking using a particle filter combined with unscented Kalman filters. Journal of Electronic Imaging 2014; 23(1):013029.
- [6] S. Frintrop and P. Jensfelt. Attentional Landmarks and Active Gaze Control for Visual SLAM. IEEE Transactions on Robotics (T-RO) 2008; 24:1054-1065.
- [7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern.
- [8] Yu, Kuan-Ting, and Li-Chen Fu. An Integrated Feature Selection Strategy for Monocular Slam 2011.
- [9] Andrea Bonarini, Wolfram Burgard, Giulio Fontana, Matteo Matteucci, Domenico Sorrenti and Juan Domingo Tardos. RAWSEEDS: Robotics Advancement through Web-publishing of Sensorial and Elaborated Extensive Data Sets. In proceedings of IROS'06 Workshop on Benchmarks in Robotics Research .....2006.
- [10] S. Frintrop. VOCUS: A visual attention system for object detection and goal-directed search. Ph.D. dissertation. Universitat Bonn, Germany, 2005, ser. LNAI. Springer; 2006.
- [11] E. Rosten and T. Drummond. Fusing Points and Lines for High Performance Tracking. Machine learning for high-speed corner detection, ECCV 2006.