Proceedings of the 2010 IEEE
International Conference on Information and Automation
June 20 - 23, Harbin, China

Monocular Vision SLAM Based on Key Feature Points Selection

Eryong Wu, Likun Zhao, Yiping Guo, Wenhui Zhou
[1]*Department of Computer Science & Technology*
Hangzhou Dianzi University, Hangzhou 310018, China
wueryong@hdu.edu.cn; zliiii@163.com;zhouwenhui@hdu.edu.cn

Qicong Wang
Department of Computer Science
Xiamen University
Xiamen, 361005, China
qcwang@xmu.edu.cn

***Abstract*— Simultaneous localization and mapping (SLAM) is an key research content of robot autonomous navigation, the visual monocular SLAM based on Extend Kalman Filter(EKF) is one important method to handle this problem. But due to high computational complexity, it has strict limits on the number and stability of the feature points, traditional method selects few corners like or straight lines as feature points, and these methods limit the application scope of EKF-SLAM. This paper proposes a key points selection method based on SIFT(Scale-invariant feature transform) feature point, on the assumption of relative uniform of the feature points' distribution, through controlling the total number of feature points effectively, the applied restriction of the visual monocular EKF-SLAM is reduced. Experiments show that this feature point selection method has a high stability for different scenes, and improves the convergence velocity.**

***Index Terms*— Robot; EKF-SLAM; Key point selection; Monocular vision; SIFT**

I. Introduction

SLAM is an important research area of the autonomous robot navigation, which aims to estimate the robot trajectory and the location of the landmarks simultaneously. However-localization requires precise landmarks locations information, while the determination of exact landmark positions need localization information too, which leading SLAM into a mutual exclusion problem[1]. If only concerned about algorithmthere are two methods to solve SLAM problem mainly: EKF based and particle filter based[2][3], while EKF based SLAM works out effectively and has been widely used in indoor environment[4]. In the sensing mode, the visual sensing, due to its advantages of easily extracting feature points, huge information and low cost, is gradually getting more attention and becoming mainstream of the SLAM research. Davison [5]has done a great deal of fruitful research, in which Inverse Depth creates a new depth estimation method , which has obvious advantages compared to the tradition ways. The computational complexity of EKF estimation is O (N^2), where N is the estimated state dimension. But for all-state EKF-SLAM estimation, the number of feature points is strictly limited, because of the features are accumulated continuously, resulting dimensional disaster problem [6]. In the past for the structured environment, image corners like Harris or straight lines are mostly taken as characteristic landmark[8][9], which achieves a better experiment result because it's fewer and relatively evenly distributed. But for general environment it does not have the above mentioned characteristics, stable visual feature extraction becomes a uneasy handled issue. Since SIFT feature points is insensitive to luminance noise, and has the capability of rotation, translation, scaling, and affine invariance, and matching correctness is very high, so it is very suitable to be landmark feature. However, hundreds of feature points extracted by SIFT method can not meet the requirements of number limitation of state dimension for EKF-SLAM approach. To solve this problem, this paper aims at, under the promise that feature points are evenly distributed, randomly selecting a certain number of feature points depending on probability requirement. This not only guarantees the number of filters required for convergence, but also improving the convergence speed of the algorithm. Finally, this paper develops a monocular vision EKF-SLAM approach which has a slight increment in accuracy and a broader scope of application. It is based on the above method of selecting feature points randomly, combined with the characteristic representation of Inverse Depth, and integrated into the camera distortion correction model.

II. Selection Method for Random Feature Points

In order to ensure the convergence of EKF-SLAM algorithm, feature points are desired to be evenly distributed within the vision field and can be observed for many times during the subsequent frames. At the same time the overall number of selected feature points must be less than a certain threshold to ensure its implementation feasibility. For the above considerations, this paper, firstly calculates feature points' distribution within the local region, then finds the significant regions of feature distribution density among the entire image, finally randomly chooses feature points according to the significance degree. When the numbers of matched feature points between observation and landmark map is less than a certain threshold, random selection is made in the

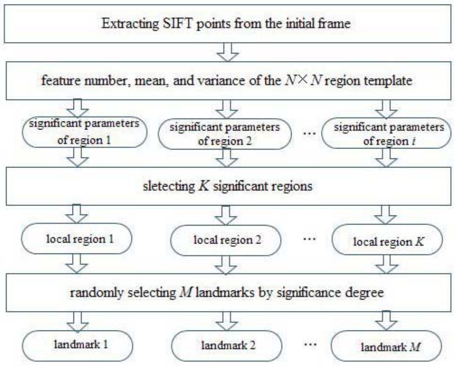
*This work is partially supported by Nature Science Foundation of Zhejiang Province(No. Y1080883 and No. Y1080967) and Supported by the open project of Zhejiang Provincial Key Laboratory of Information Network Technology(No. 200808 and No. 200807)

978-1-4244-5704-5/10/$26.00 ©2010 IEEE

unmatched significance region just as the above mentioned method; otherwise no new feature point will be selected. Specific methods can be summarized in the following two steps:

Step 1. Initial landmarks selection

Extracting SIFT feature points from video, and then obtain the key frame as initial frame, which has relatively stable key points number and frame-to-frame matching; dividing the image into $N \times N$ regions, calculate the number of feature points in each region and feature locations' mean and variance; choose K significant local feature regions with high significance; finally selecting M landmarks in accordance with the significance degree(the number of feature points $\times$ feature locations' variance). Following is the diagram of initial landmarks selection process:



Extracting SIFT points from the initial frame
feature number, mean, and variance of the $N\times N$ region template
significant parameters of region 1
significant parameters of region 2
...
significant parameters of region i
sletecting K significant regions
local region 1
local region 2
...
local region K
randomly selecting M landmarks by significance degree
landmark 1
landmark 2
...
landmark M

Fig. 1. The diagram of initial landmarks selection

Step 2. Landmark Augmentation

Matching current frame's SIFT features with landmark map, if the matched number is less a threshold, then do landmark augmentation. The methods is: extracting the feature points' significance region of the current frame; selecting the regions that not include landmarks map feature, then depending on the degree of these regions, randomly selecting L feature points as new landmarks; meanwhile the distance between the selected features of the same area should be greater than a certain threshold. When the number of the landmarks is greater than threshold T, no more augmentation needed. Fig. 2 shows the augmentation process.

III. Visual Monocular EKF-SLAM Based on Inverse Depth Method

A. Motion Model

The camera state describes as $\boldsymbol{r}^{wc} = (x, y, z)^T$ position and quaternion $\boldsymbol{q}^{wc} = (q_w, q_x, q_y, q_z)^T$ relative to cartesian coordinate system, denoted as $\boldsymbol{x}_v$:

$$\boldsymbol{x}_v = (\boldsymbol{r}^{wc}; \boldsymbol{q}^{wc}) \tag{1}$$

where superscript wc means camera relative to world coordinate.

Then the camera's movement is modeled as linear velocity model, the frame-to-frame orientation change $\boldsymbol{q}_0$ and translation$\boldsymbol{t}_0$ can be got from multiview geometry such as paper [10]. Camera's state evolution is a process of rigid Euclidean transformation. According the above velocity change model, the state update equation for the camera is:

$$\boldsymbol{f}_v = \begin{pmatrix} \boldsymbol{r}^{wc}_{k+1} \\ \boldsymbol{q}^{wc}_{k+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}^{wc}_{k} + \boldsymbol{t}_0 \\ \boldsymbol{q}^{wc}_{k} \times \boldsymbol{q}_0 \end{pmatrix} \tag{2}$$

The camera state variance is $\boldsymbol{P}_v(k \mid k)$ in time step$t$, and then the predicted variance in time t+1 is:

$$\boldsymbol{P}_v(k+1 \mid k) = \frac{\partial \boldsymbol{f}_v}{\partial \boldsymbol{x}_v} \boldsymbol{P}_v(k \mid k) \frac{\partial \boldsymbol{f}_v}{\partial \boldsymbol{x}_v}^T + \boldsymbol{R} \tag{3}$$

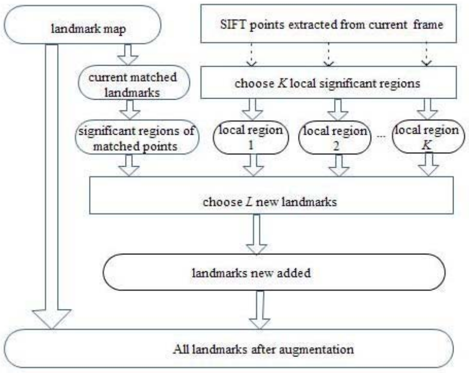
where $\boldsymbol{R}$ is the predicted variance of the random white noise.

B. Observation model

According to the Inverse Depth method of J.M.M. Montiel [5], a landmark in the map is defined by the dimension six state vector:

$$\boldsymbol{y}_i = (x_i, y_i, z_i, \theta_i, \phi_i, \rho_i)^T \tag{4}$$

Where, $(x_i, y_i, z_i)^T$ presents the camera optical center's Cartesian coordinates that this landmark is observed firstly; $\theta_i$andϕ_i respectively stands for polar angle and elevation angle in accordance with the optical center of the observation vector in the world coordinate system;ρ_i means inverse depth. Supposing the state of the observation landmark is $\boldsymbol{y}_i$, $\boldsymbol{m}(\theta_i, \phi_i)$ is the representation for the ray projection vector in the world coordinate system. The landmark observation model is shown in Fig. 3:



landmark map
SIFT points extracted from current frame
current matched landmarks
choose K local significant regions
significant regions of matched points
local region 1
local region 2
...
local region K
choose L new landmarks
landmarks new added
All landmarks after augmentation

Fig. 2. The diagram of landmarks augmentation

Then the theoretical projection of this landmark in time t+1 in the camera coordination is supposed to be:

$$\boldsymbol{z}^c = \boldsymbol{R}^{wc}((x_i, y_i, z_i)^T + \frac{1}{\rho_i}\boldsymbol{m}(\theta_i, \phi_i) - \boldsymbol{r}^{wc}) \qquad (5)$$

Where, $\boldsymbol{R}^{wc}$ is the matrix representation of the camera orientation, $\boldsymbol{r}^{wc}$ represents the Cartesian coordinates of camera optical center, superscript wc means the camera coordinates system compared to the world coordinates.

The intrinsic parameters of camera $\boldsymbol{M}$ and the len distortion parameter $\boldsymbol{M}_d$ can be obtained through the calibration process [9]:

$$\boldsymbol{M} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \qquad (6)$$

$$\boldsymbol{M}_d = \begin{pmatrix} k_1 & k_2 & k_3 & k_4 \end{pmatrix} \qquad (7)$$

With the known $\boldsymbol{z}^c$, we can get the projection coordinates after distortion $\boldsymbol{z}_d = (z_x, z_y, z_z)^T$ as the method introduced in the literature [9], and then the projected image coordinates is:

$$(uv)^T = \boldsymbol{h}(\boldsymbol{x}_v, \boldsymbol{y}_i, \boldsymbol{M}, \boldsymbol{M}_d) = \begin{pmatrix} c_x - \frac{z_x}{z_y} f_x \\ c_y - \frac{z_y}{z_z} f_y \end{pmatrix} \qquad (8)$$

Form Eq.8 we can get the variance from the observed coordinates:

$$\boldsymbol{P}_h = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_v} \boldsymbol{P}_{x_v} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_v}^T + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}_i} \boldsymbol{P}_{y_i} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}_i}^T + \boldsymbol{Q} \qquad (9)$$

Where, $\boldsymbol{Q}$ is variance from the observed white noise by feature points.

C. State Update

Both the camera state and all the landmarks location in the map should be estimated at the same time, so the all state vector is:

$$\boldsymbol{x} = (\boldsymbol{x}_v, \boldsymbol{y}_1, \boldsymbol{y}_2, ...\boldsymbol{y}_n)^T \qquad (10)$$

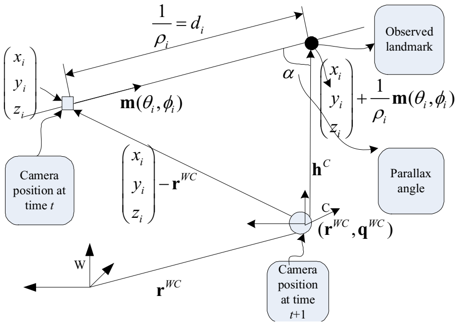

$\frac{1}{\rho_i} = d_i$
(x_i, y_i, z_i)
$\mathbf{m}(\theta_i, \phi_i)$
α
Observed landmark
$(x_i, y_i, z_i) + \frac{1}{\rho_i}\mathbf{m}(\theta_i, \phi_i)$
Camera position at time t
$(x_i, y_i, z_i) - \mathbf{r}^{WC}$
$\mathbf{h}^C$
Parallax angle
C
$(\mathbf{r}^{WC}, \mathbf{q}^{WC})$
W
$\mathbf{r}^{WC}$
Camera position at time t+1

Fig. 3. Observation model

Where, $\boldsymbol{x}_v, \boldsymbol{y}_i, (i = 1, 2, ..., n)$ are respectively camera state and absolute Coordinates of landmarks in the map.

According to the above mentioned motion model, the predicted part of the SLAM with full-state Kalman filter can be described as:

$$\boldsymbol{x}_v(k+1 \mid k) = \boldsymbol{f}_v(\boldsymbol{x}(k \mid k)) \qquad (11)$$

$$\boldsymbol{P}_v(k+1 \mid k) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}_v} \boldsymbol{P}_v(k \mid k) \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}_v}^T + \boldsymbol{R} \qquad (12)$$

whereboth the variance and mean of the predicted landmarks remain unchanged.

In time $t+1$, the theoretical image observation coordinates $\boldsymbol{z}^{'}$ and observation Jacobian matrix H can be represented as follows:

$$\boldsymbol{z}^{'} = \boldsymbol{h}(\boldsymbol{x}(k+1 \mid k), \boldsymbol{y}, \boldsymbol{M}, \boldsymbol{M}_d) \qquad (13)$$

$$\boldsymbol{H} = (\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}_v}, \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}_1}, \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}_2}, ..., \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{y}_n}) \qquad (14)$$

Then the Kalman gain coefficient $\boldsymbol{K}$ is:

$$\boldsymbol{P}(k+1 \mid k)\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}(k+1 \mid k)\boldsymbol{H}^T) + diag(\boldsymbol{R}, \boldsymbol{Q}, ..., \boldsymbol{Q}) \qquad (15)$$

So the all state update process is:

$$(x)(k+1 \mid k+1) = (x)(k+1 \mid k) + \boldsymbol{K}(\boldsymbol{z} - \boldsymbol{z}^{'}) \qquad (16)$$

$$(P)(k+1 \mid k+1) = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})(P)(k+1 \mid k) \qquad (17)$$

Where$\boldsymbol{z}$ is actual image observation coordinates of matched landmarks, $\boldsymbol{I}$ stands for the unit matrix, $\boldsymbol{Q}$ represents the observation noise.

If there is any new landmark in, then depending on the current camera state and image observation coordinates, with the method shown in Fig. 3, get new landmarks state and variance, augment them to the all state estimation vector $\boldsymbol{x}(k+1 \mid k+1)$ and $\boldsymbol{P}(k+1 \mid k+1)$.

IV. Experiments

This experiment uses Gsou Q5-V network camera with a resolution of 320 × 240, 25 frames per second; selects four different scenes, taking 3-5 seconds shooting to each scene with hand-held camera, and on-line calculating the collected data through the USB interface. Computing platforms: ThinkPad SL400, CPU for the Core 2 Duo 2.1G, memory 2G, MATLAB7.0 as the software development language. Finally, the result has been saved for offline analysis.

A. Experiments and analysis of the extracted feature points

For the four different scenes in the Fig.4, extracting SIFT feature points from the initial frame, dividing the image evenly to 8×6 regions, calculating the distribution parameters in each region, selecting 12 local regions by significance degree, finally evenly collecting 10 feature points depending on the significance information. The result is shown in Fig. 5

If the matched number between the current frame and map landmarks is less than 4, then do landmark augmentation. Firstly, obtaining the significance regions of the unmatched map landmarks, then choosing 3 feature points as new landmarks according to the significance degree, but the distance of different features should be bigger than 20 pixels.

Calculate the distance variance of the current frame and new added image features, which is taken as the judgment of choosing feature points distribution. The method for calculating distance is:

$$p_{uv} = \frac{1}{n}\sum_{i=1}^{n}((u-\bar{u})^2-(v-\bar{v})^2)/(WH) \qquad (18)$$

Where, $WandH$ respectively stands for the width and height of the image, n is the number of the current observation landmarks.

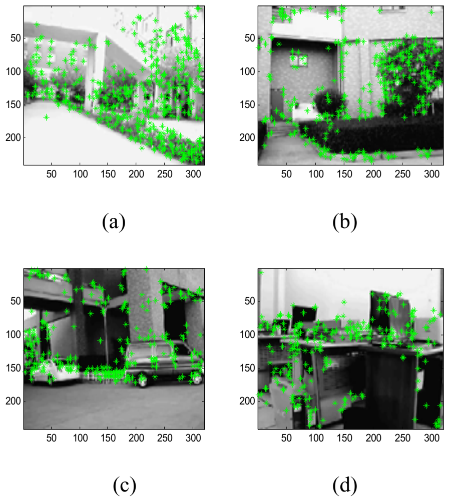
The distribution diagram of landmarks distance variance is shown in Fig. 6, from the Figure we can see that, with the landmarks distance variance growing larger, the distribution is inclined to be stable. Just as in Fig. 6(c), about at 60th frame, because of the fewer matches of the map landmarks, the variance has a sharply declining. However, by the landmarks augmentation introduced in this paper, it cans swiftly callback the variance distribution, which reflects the validity of using this method to extracting feature points.

B. Experiments and analysis of SLAM

Firstly we use the method mentioned in the literature [9] to attain the inner parameters of camera and distortion parameters, then we choose image Fig.4(d) as experiment scene, applying the features extracted method introduced in the above A, conducting the monocular SLAM experiment online.

After doing EKF filter in 10 frames, Fig.7 shows that the location of landmarks has converged to a certain level, which indirectly reflecting that adding distortion parameters correction will improve the convergence speed. The left image in Fig.7 shows all the current landmarks location in the frame, the blue dots in the right part of Fig.7 stands for the two-dimensional camera path, red oval on behalf of landmarks variance in the x and z direction.

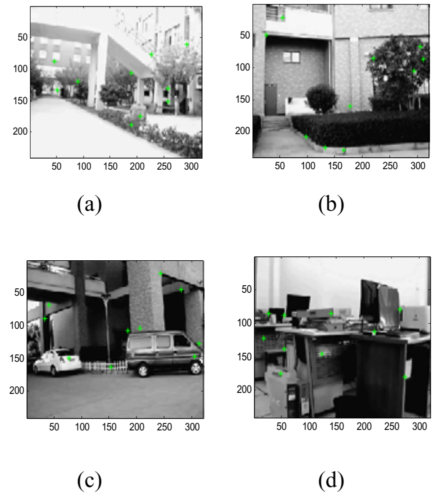
Fig.8 indicates the landmarks convergence of the last frame, from the left picture we can see 34 landmarks in the final image, the right one shows us red ovals, which has larger area, corresponding to the landmarks added into landmark map recently. Because the number of filtering is fewer, which resulting in a larger landmarks variance, incomplete convergence, and un-obvious distance information. Fig.9 shows convergence of points of nearer sights, from which we can see obvious distinction of points from far and near sights, and also a steady convergence result. Camera's three-dimensional trajectory is shown in Fig. 10. Fig.7 to Fig.9 only shows two-dimensional results of the landmarks in order to be intuitive.



(a) (b)

(c) (d)

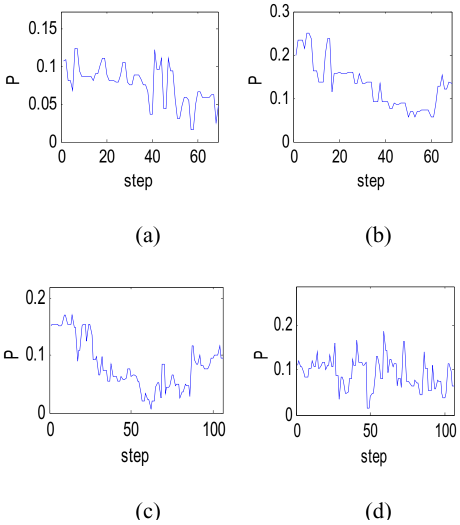
Fig. 4. Initial frame's SIFT points distributions for different scenes



(a) (b)

(c) (d)

Fig. 5. Initial frame's extracted landmarks distribution for different scenes



(a) (b)

(c) (d)

Fig. 6. Distance variance distribution of landmarks for different scenes

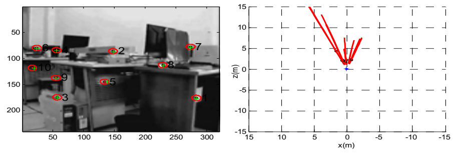


Fig. 7. Convergence result of step 10

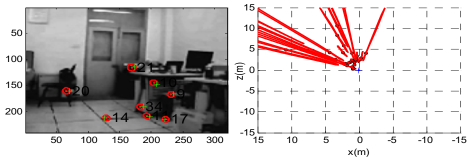


Fig. 8. Convergence result of step 100

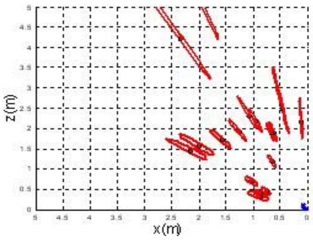


Fig. 9. Partial magnified convergence result of step 100

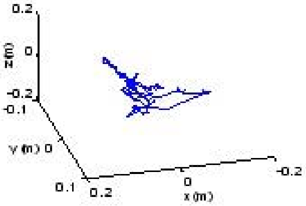


Fig. 10. Camera three dimensional trajectory

V. Conclusion

Based on traditional EKF-SLAM, and using Inverse Depth landmark model, one Monocular vision SLAM method with key feature points selection is advanced. The feature landmarks selected by this method has the advantages of ensuring landmarks distributed relatively evenly and the total number of feature selected decreased obviously. Experiments show the algorithm improves the convergence speed and gets a better result finally.

References

[1] J. J. Leanard, H. F. Durrant-Whyte, ”Mobile robot localization by tracking geometric beacons”, IEEE Trans. on Robotics and Autoumation, Vol. 7, No. 3, pp. 376-382, 1991.

[2] G Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba,”A solution to the simultaneous localization and map building (SLAM) problem”, IEEE Trans. on Robotics and Automation, Vol. 17, No. 3, pp. 229-241, 2001.

[3] K. Murphy, ”Bayesian map learning in dynamic environments”, In Advances in Neural Information Processing Systems (NIPS), pp.1015-1021, 1999.

[4] J. Civera, A. J. Davison, and J. M. M. Montiel, ”Interacting multiple model monocular SLAM”, In Proceedings of the IEEE International Conference on Robotics and Automation, pp. 3704 - 3709, May, 2008.

[5] A. D. J. Montiel, J. Civera, ”Unified inverse depth parametrization for monocular slam,” in Proceedings of Robotics: Science and Systems, Philadelphia, PA, June 2006.

[6] J. Guivant and E. Nebot,”Improving computational and memory requirements of simultaneous localization and map building algorithms”, IEEE International Conference on Robotics and Automation, pp. 2731-2736, 2002.

[7] E. Eade and T. Drummond, ”Unified loop closing and recovery for real time monocular slam”, In Proceedings of the British Machine Vision Conference (BMVC), 2008.

[8] E. Eade and T. Drummond, ”Monocular SLAM as a graph of coalesced observations”, In Proceedings of the International Conference on Computer Vision (ICCV), 2007.

[9] Jean-Yves Bouguet. Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/ bouguetj/calib doc/index.html.

[10] Eryong Wu, Wenhui Zhou, Guojun Dai, Qicong Wang, ”Monocular vision SLAM for large scale outdoor environment”, International Conference on Mechatronics and Automation, p.2037-2041, 2009.