Localization based on vision using stereo System and SIFT features

Somaia Mohamed∗1, Aboelmagd Noureldin Mail2, Mohamed Hesham El-Sayed1, Magdi Fekri Ragaey1

1Faculty of Engineering, Cairo University, Cairo, Egypt
2School of Engineering, Queen’s University, Ontario, Canada

Received: October 13, 2015 Accepted: October 30, 2015 Online Published: November 16, 2015
DOI: 10.5430/air.v5n1p95 URL: http://dx.doi.org/10.5430/air.v5n1p95

ABSTRACT

The traditional method uses the GPS signal to determine the location of a vehicle or Mobile Robot, but these methods sometimes fail to determine the exact position of the object especially in urban areas, or inside buildings. It fails to determine the position for multiple reasons such as the multi-fading path, obstacles, rain, and snow which all have an effect on the signal of the GPS. The developed algorithm is intended to get the location of the moving objects by using the computer vision system. The results of the proposed algorithm give higher performance (around 33 second and with a resolution of 10 meters) compared to the other systems on the long trajectory. We use the single camera system to select the suitable features and the stereo camera systems to obtain the locations of the object. It is therefore different from the other methods that are dependent on maps to compare the image features with the features of the map to determine the location of the object.

Key Words: Computer vision, Ego-motion, Localization, Vehicle, Navigation, Object tracking, Object detection, Robot, Visual odometry

1. INTRODUCTION

The navigation system nowadays tries to process the error of the convention all systems or improve the performance of conventional systems by using inertial sensors, LIDAR and computer vision with the GPS system. The Imaging technology has achieved great progress in the last decade. Cameras have become cheaper, smaller and higher in quality than before. Additionally, computing power attracts the attention of the researchers and the computing platforms are geared towards parallelization. Therefore, the advances in the hardware are reflected in the computer vision, which enables it to achieve great progress in vehicle localization and make the researchers more attracted towards this research area.[1] There are many methods in computer vision that were used to solve the navigation problems. Xu et al. use the vision and curvature estimates to enable localization on a network road. The average error in downtown was from 3.7 m to 15.87 m and the average error on a parallel road was from 8.79 m to 17.3 m.[2] The method that was presented in Ref.,[3] uses a combination of the metric and topological localization to achieve the advantages of the two methods. They extracted the features of the route and the location of each feature by the GPS system to build the database. Then they collected the images, extracted SURF features for the same route, and compared these features with the database to get the location of the vehicle. The average error of this method was between 2.7 m and 10 m.[3]

Pink et al. demonstrate a method for vehicle pose estima-
tion and motion tracking. The algorithm used some ideas from the visual odometry research and the map features, which were extracted automatically from aerial images. The accumulated errors of the algorithm after 2 seconds and 50 frames was 0.56 m. It depended on the storage data from maps. The dependence on data from maps as in Refs. put some limitations on the system because sometimes these maps may not be available on global systems. Stein et al. used the values from two consecutive images and combined it in a global probability function. This combination enabled the algorithm to ignore a large number of outlier points. All processing was done in offline mode so it will not be suitable for the real time applications. Additionally, the actual road of the experiment was between 63 m and 100 m, as well as it did not have an accurate ground truth table for the traveled distance of the trajectory.

This paper will demonstrate an algorithm to determine the location of a vehicle based on computer vision. The related work will be explained in Section 1. Section 2 will demonstrate the proposed algorithm. Section 3 will show the experiments and results. Finally, the conclusion will be in Section 4.

2. METHOD

The main components of any vision localization system are the cameras that are used to collect the data and the calibration process of these cameras according to the system single or stereo, the features that are extracted from the collected data, and the methodology that will be used to determine the position of the target.

2.1 The cameras

The single camera system primarily assists the system of localization to estimate the location and to predict the vehicle position on the image plane. The stereo camera system is any system containing two or more cameras, which are used mainly in the localization to obtain the location of the object in a metric scale. A single camera is shown in Figure 1(a) and a stereo camera is shown in Figure 1(b).

![Figure 1](image1.png)

**Figure 1.** Pinhole camera model

The first step in using any camera is the calibration process. The calibration of the cameras is necessary to extract the intrinsic parameters (focal length, principal point, skew coefficient and distortions) and extrinsic parameters such as transformation (R) and rotation (T) matrix. The relationship between the point in world coordinate and the point in the image plane is shown in Figure 2. We can transfer the point from world coordinate (Rw) to camera frame (Rc) as shown on the graph in step 1 by Equation 1.

\[
\begin{bmatrix}
Xc \\
Yc \\
Zc \\
1
\end{bmatrix} = [R]
\begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix} + t = [R]
\begin{bmatrix}
tx \\
ty \\
tz
\end{bmatrix} = [T]
\begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix}
\]

(1)

The transformation from Rc to image plane (Rr) as in step 2 will be according to Equation 2.

\[
\begin{bmatrix}
x \\
y \\
z
\end{bmatrix} = [P]
\begin{bmatrix}
Xc \\
Yc \\
Zc \\
1
\end{bmatrix}
\]

(2)

![Figure 2](image2.png)

**Figure 2.** The relation between the world frame and image frame

The third transformation from the image plane to the pixel plane or sensor reference frame is shown on Figure 2 in step 3 by the following equation:

\[
\begin{bmatrix}
u \\
v \\
1
\end{bmatrix} = 
\begin{bmatrix}
kx & 0 & cx + cy cot \theta \\
0 & kx cot \theta & cy sin \theta \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix} = [A]
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

(3)
In the beginning, we used the single camera system to draw the trajectory path to compare the results with the shape of trajectory from the ground truth table, which was extracted from the GPS camera file. This part helped with the choice of features. Then, the work is completed by using the stereo system to get the location of the vehicle in 3D coordinates.

2.2 Features extraction

Most visual localization methods are interested in the extraction of local features from the images. Feature selection is one of the important steps in a vision system. There are many parameters, which control the process of selection, such as the processing time, accuracy, affine transformation, scale changes, illumination changes, and blurring. The survey in Refs. [1, 9] found that the SIFT features (Scale Invariant Features Transform) and SURF (Speeded Up Robust Features) are robust local features detector and suitable for the applications of tracking and localization. The next subsections give brief description about SIFT and SURF features.

Badino et al. [3] use the SURF features for localization. They used high-resolution panoramic images collected over long periods. The performance of SURF features is better than other local features for outdoor localization. Also, A. Ascani et al. [11] extracted SIFT and SURF features from the images which were collected in different environmental conditions such as the change in light during the day, and in different days, both indoors and outdoors, for topological and metric localization. The performance of group matching features in indoor localization for topological and metric localization is better when using SIFT features than when using SURF features.

According to the comparison between SURF and SIFT in Refs. [11, 12] it is found that the performance of SIFT and SURF is similar, but SIFT is more stable than SURF in the rotation and illumination changes.

2.2.1 SIFT features

The SIFT algorithm is extracted features that are invariant to rotation, scaling, illumination and affine transformation of images to perform matching of different views of an object or scene. Steps for extracting SIFT features are as follows:[13, 14]

1. Scale-space extrema detection is based on difference-of-Gaussian function \( D(x, y, \sigma) \) to identify keypoints locations and scales that can be frequently assigned under differing views of the same object. The scale space of an image is defined as a function \( L(x, y, \sigma) \) that is obtained from the convolution of a variable scale Gaussian function \( G(x, y, \sigma) \) with an input image \( I(x, y) \) according to Equations 4, 5 and 6.[15, 16]

\[
L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)
\]
G(x, y, σ) = \frac{1}{2\pi\sigma} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (5)

D(x, y, σ) = L(x, y, k\sigma) - L(x, y, σ) \quad (6)

(2) Keypoint localization: Select keypoints according to measures of their stability.\[15\]

(3) Orientation assignment: One or more orientation is assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, therefore providing invariance to these transformations.\[15\]

(4) Keypoint descriptor: The local image gradients are measured at the selected scale in the region around each keypoint. These are transformed into a representation that allows for significant levels of local shape distortion and change in illumination.\[15\]

Extreme values are detected at the different scales of the image, and are the keypoint candidates. Secondly, the Taylor series and Hessian matrix are used to determine stable keypoints. Thirdly, the gradient orientation is assigned to the keypoint by using its neighboring pixels, and finally, keypoint descriptor is obtained.\[13\]

The extraction of SIFT features in this work is applied on frames with size around 311 KB and dimensions 1,920 × 1,080. Figure 5 shows the extracted matching features from frame \( n \) and \( n + 1 \) and after processing to discard the outlier points.

Figure 5. Samples for the SIFT features matching from frame \( n \) and \( n + 1 \)

**2.2.2 SURF features**

SURF is a novel scale and rotation invariant interest point detector, descriptor and matching. It is extracted by relying on integral images for image convolution. Extraction of SURF features is divided into three steps:

(1) Interest point detection is selected at distinctive locations in the image, such as corners, blobs, and T-junctions. It is based on the Hessian matrix that approximates second order Gaussian derivative with box filters by using integral images.\[17\] The hessian matrix has good performance in computation time and performance. The hessian matrix \( H(x, \sigma) \) of the point \( x = (x, y) \) in image \( I \) and scale \( \sigma \) is calculated according to Equation 7.\[17, 18\]

\[
H(x, \sigma) = \begin{bmatrix}
L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\
L_{xy}(x, \sigma) & L_{yy}(x, \sigma)
\end{bmatrix}
\]  

(7)
$L_{xx}$ is the convolution of the second order derivative of Gaussian with the image $I$ in the point $x$.

(2) Orientation assignment that is determined by constructing a circular region with radius $6\sigma$ around the detected interest point, with $\sigma$ the scale at which the interest point is detected and the dominant orientation describes the orientation of interest point.$^{[18,19]}$

(3) Interest point descriptor is represented by a feature vector that is constructed by extracting a square window around the interest point and computing the Haar wavelet responses in horizontal and vertical directions.$^{[16]}$ It also has to be distinctive and at the same time, robust to noise, detection error, and scale and illumination.$^{[17,18]}$

2.3 The proposed algorithm

The proposed algorithm is intended to determine the location of moving objects (i.e. any vehicle or robot carrying a vision system) by vision system as shown in Figure 6 and described in Algorithm 1:

**Algorithm 1 Proposed algorithm**

Read Camera’s frames $F_{Li}$, $F_{Ri}$, $F_{Li+1}$, $F_{Ri+1}$ and initial position $P_{old}$

Compute features $D_{Li}$, $D_{Li+1}$, $D_{Ri}$, $D_{Ri+1}$ of frame $i$ and $i+1$.

Calculate matching ($M_1$) between $D_{Li}$ and $D_{Li+1}$.

If $M_1 > 0$

Calculate matching ($M_2$) between $D_{L1}$, $D_{R1}$, $D_{L2}$, $D_{R2}$.

If $M_2 > 0$

Get matching points $mP_{L1}$, $mP_{R1}$, $mP_{L2}$, $mP_{R2}$.

Get the trust matching points by RANSAC algorithm.

Get the $P_{3dL1}$, $P_{3dL2}$ in 3-dimensions metric.

Get $DP_L = P_{3dL1} - P_{3dL2}$.

$P_{new} = P_{old} + DP_L$.

Else

Get new frames.

End

Else

Get new frames.

End

Update the position of vehicle $P_{old} = P_{new}$.

3. RESULTS

In our experiment, the dataset was generated by using two vision system, single camera system and stereo system on two different barriers (vehicle and robot as shown in Figures 12 and 13). The main hardware components used in the vision system were the cameras, the vehicle for outdoor experiments, and the robot for indoor experiments. The cameras used were MiVue 358 with the following specifications: 1,080 p Full HD resolution ($1,920 \times 1,080$ pixels), 30 fps. The explanation of the experiment set up will be described in the following subsections.

3.1 Experiment 1

This experiment compares the output of the purposed algorithm when it uses SURF features with the output of ground truth table. The output of this experiment is shown in Figure 7, the output trajectory according to SURF features is diverged from the beginning of the trajectory compared with the output trajectory of ground truth table.

3.2 Experiment 2

The single camera system was used to draw the trajectory path and compare the output path from the vision algorithm with the output path from the GPS Camera File. This experiment was carried out by using the dataset collected from the outdoor localization by a vehicle in downtown Kingston as shown in Figure 13. The time of trajectory was around 161 second and the average speed of the vehicle was around 18 km/h. The output of the trajectory by the vision system is shown in Figure 8(a) and the output from the ground truth table is shown in Figure 8(b).
3.3 Experiment 3
We repeated experiment 1 here but on a different trajectory. The average speed of this trajectory was 18 km/h and the time length was 157 seconds. The output of the experiment by the vision system is shown in Figure 9(a) and the output from the GPS file or ground truth table is shown in the Figure 9(b).

The output graph of experiment 1 and 2 shows that the route of the trajectory is similar to the output from the ground truth table. This proves that the SIFT features give accurate results in the image plane with the outdoor localization. This step therefore helps with the next step to obtain the position of the vehicle in the 3D dimension.

3.4 Experiment 4
In this experiment, the proposed algorithm will determine the position of the vehicle by the stereo system as shown in Figure 13 and compare it with the positions from the GPS camera file. The algorithm gives an accuracy of around 10 m for 33 seconds from the trajectory as shown in Figure 10.

3.5 Experiment 5
This experiment is the same as experiment 3 but using different data. The average speed was 10 km/h and the time length was 66 sec. The output of the experiment is shown in Figure 11.
4. DISCUSSION

SIFT features are powerful features for the outdoor localization according to the results from experiments 1, 3 and 5. The proposed algorithm will be suitable with the trajectory of less than 33 seconds because if it is longer than this, the output will diverge. To improve the proposed algorithm and to solve the divergence problem with the long trajectory, we will try to get methods to reject the outlier points such as RANSAC (RANdom SAmple Consensus) or Hough transform.
REFERENCES


