Visual Odometry using Stereo Vision

Rashmi K V

***Abstract*— Conventional non-vision based navigation systems relying on purely Global Positioning System (GPS) or inertial sensors can provide the 3D position or orientation of the user. However GPS is often not available in forested regions and cannot be used indoors. Visual odometry provides an independent method to estimate position and orientation of the user/system based on the images captured by the moving user accurately. Vision based systems also provide information (e.g. images, 3D location of landmarks, detection of scene objects) about the scene that the user is looking at. In this project, a set of techniques are used for the accurate pose and position estimation of the moving vehicle for autonomous navigation using the images obtained from two cameras placed at two different locations of the same area on the top of the vehicle. These cases are referred to as stereo vision. Stereo vision provides a method for the 3D reconstruction of the environment which is required for pose and position estimation. Firstly, a set of images are captured. The Harris corner detector is utilized to automatically extract a set of feature points from the images and then feature matching is done using correlation based matching. Triangulation is applied on feature points to find the 3D co-ordinates. Next, a new set of images is captured. Then repeat the same technique for the new set of images too. Finally, by using the 3D feature points, obtained from the first set of images and the new set of images, the pose and position estimation of moving vehicle is done using QUEST algorithm.**

***Keywords*— Stereo, Rover, Odometry, Rectification, Quaternion.**

I. INTRODUCTION

Robot navigation is a well-known problem. Efforts have been on for several years for obtaining information about the environment in which the robot is required to operate to enable it to move through it. Robots, provided with mounted cameras and movement capabilities, can be used for various purposes. The methods given in this report can be used to estimate the pose and position of a robot or an autonomous vehicle. This method can be used as an integral part of the path planning, finding the path traced by rovers and navigation of the robot or the autonomous vehicle [1].

Odometry is the study of position estimation during wheeled vehicle navigation. The term is also sometimes used to describe the distance travelled by a wheeled vehicle. Visual odometry is the process of determining the position and orientation of an object by analyzing the sequence of images or video signal. It allows for enhanced navigational accuracy in robots or vehicles using any type of locomotion on any surface. it can independently estimate the position and 3D orientation accurately by using only image streams captured from two or more cameras. The images can be obtained using two cameras, or sequentially, using a moving camera. These cases are referred to as stereo vision. Stereo vision refers to the ability to infer information on the 3D structure and distance of a scene from two or more images taken from different viewpoints.

A. Computer vision

Computer vision is the science and technology of machines that the machine is able to extract information from an image that is necessary to solve some task. Computer vision differs from Image processing in that image processing mainly concerns image to image properties whereas the main target of computer vision is the 3D world. In order to estimate the pose and position of moving user the camera must be calibrated. This is equivalent to assuming that the camera intrinsic and extrinsic parameters are known. The extrinsic parameters are the parameters that define the location and orientation of the camera reference frame with respect to known world reference frame. The intrinsic parameters are the parameters necessary to link to the pixel co-ordinates of an image point with the corresponding co-ordinates in the camera reference frame.

1) Intrinsic parameters

The camera intrinsic parameters are defined as the focal length, f, the location of the image center in pixel co-ordinates (o_x, o_y) the effective pixel size in the horizontal and vertical direction (s_x, s_y).

Let x and y be the pixel co-ordinates of the feature point x_{im} and y_{im} also and are the co-ordinates of the feature point in the image plane. Let X_c, Y_c and Z_c are the co-ordinates in the camera plane. Then the relation between pixels co-ordinates and the co-ordinates of the image plane is given by eq. (1).

$$x_{im}=-(x-o_x)s_x$$
$$y_{im}=-(y-o_y)s_y \quad (1)$$

The relation between image plane and camera plane is given by

$$X_C = Z_C \frac{x_{im}}{f}$$
$$Y_C = Z_C \frac{y_{im}}{f} \quad (2)$$

If there are no rotation in any axis then the camera reference frame is itself is the world reference frame. If there is some rotation then the co-ordinates are need to be find using camera extrinsic parameters [2] [3].

2) Extrinsic parameters

The camera extrinsic parameters are the Translation vector, T, and the Rotation matrix, R, which specify the transformation between the camera and the world reference frame [2] [3].

Computer vision algorithms for reconstructing the 3D structure of a scene or computing the position of objects in space need equations linking the co-ordinates of points of 3D space with the co-ordinates of their corresponding image points. These equations are written in the camera reference frame, but it is assumed that

- The camera reference frame can be located with respect to some other, known, reference frame.
- The co-ordinates of the image points in the camera reference frame can be obtained from pixel co-ordinates, the only ones directly available from the image.

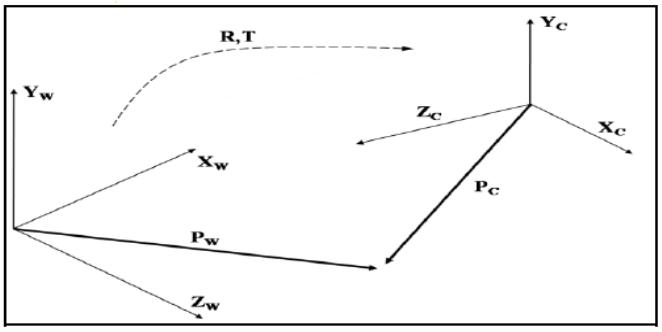

Fig 1: The relation between camera and world coordinate frames

The relation between the co-ordinates of a point P in world and camera frame P_W and P_C is

$$P_C = R(P_W - T) \quad (3)$$

With

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \quad (4)$$

B. Stereo vision

Stereo means reconstructing 3D effect using two or more 1D/2D signals. Stereo vision uses two or more cameras, but the condition is that both the cameras must be coplanar and parallel to each other.

Triangulation

The technique for gauging depth information given two offset images is called triangulation. Triangulation makes use of the terms, image planes, optical axes, disparity, baseline, focal length[14]. The following examples show how the triangulation technique works. The optical axes of the cameras are aligned parallel and separated by a baseline of distance, b. A coordinate system is attached in which the x-axis is parallel to the baseline and the z-axis is parallel to the optical axes. The points labelled "Left Camera" and "Right Camera" is the focal points of two cameras. The distance is the perpendicular distance from each focal point to its corresponding image plane. Point P is some point in space which appears in the images taken by these cameras. Point P has co-ordinates (x, y, z) measured with respect to a reference frame that is fixed to the two cameras and whose origin is at the midpoint of the line connecting the focal points. The projection of point P is shown as P_r in the right image and P_1 in the left image and the co-ordinates of these points are written as (x_r, y_r) and (x_1, y_2) in terms of the image plane coordinate systems and for tracing the path of a rover or a robot using stereo vision requires a geometrical setup, the front view of the geometrical setup is shown in the Fig. 2.

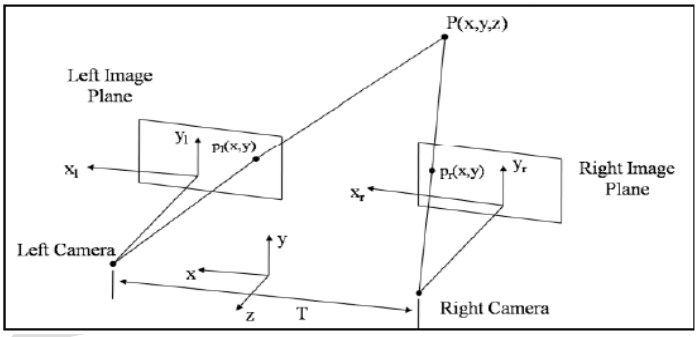

Fig 2: The geometry of stereo vision

The optical axes of the cameras are aligned parallel and separated by a baseline of distance, b. A coordinate system is attached in which the x-axis is parallel to the baseline and the z-axis is parallel to the optical axes. The points labelled "Left camera" and "Right Camera" is the focal points of two cameras. The distance is the perpendicular distance from each focal point to its corresponding image plane. Point P is some point in space which appears in the images taken by these cameras. Point P has co-ordinates (x, y, z) measured with respect to a reference frame that is fixed to the two cameras and whose origin is at the midpoint of the line connecting the focal points. The projection of point P is shown as P_r in the right image and P_1 in the left image and the co-ordinates of these points are written as (x_r, y_r) and (x_1, y_2) in terms of the image plane coordinate systems. If the cameras are not parallel then epopolar rectification is done for stereo images [4] [5].

II. VISUAL ODOMETRY USING STEREO VISION

The block diagram for the proposed work is shown in Fig. 1.1. It requires two cameras which are of the same type having same specifications (focal length, sensor width, sensor height) [1].

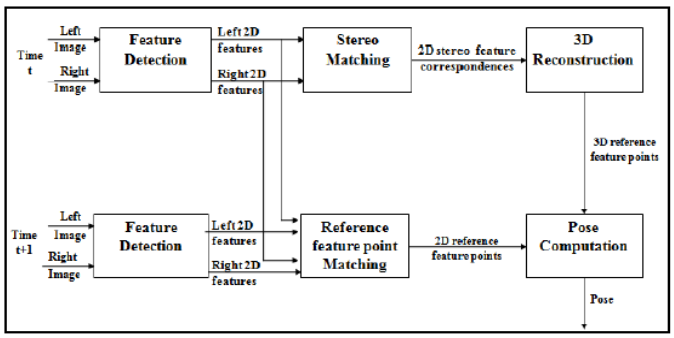

Fig 3: Block diagram for visual odometry

Firstly two cameras are mounted are such that the two cameras are coplanar and parallel to each other separated by some distance called baseline. Given a pair of stereo images, our system starts with detecting a set of potential feature points for pose estimation. The Harris corner detector is utilized to automatically extract a set of feature points from the left and right images respectively [6]. A stereo matching algorithm is used to find correspondences between the extracted feature points between left and right images . The 3D coordinates of each feature point are obtained by triangulation using the feature correspondences. These obtained 3D feature points serve as the reference points for pose computation when a new pair of stereo images arrives. Next, a new set of images is captured. Then repeat the same technique for the new set of images too. Finally, by using the 3D feature points, obtained from the first set of images and the new set of images, the pose and position estimation of moving vehicle is done using QUEST algorithm [7] [8].

III. EXPERIMENTAL RESULTS

The entire work has been carried out using MATLAB 7.10.0 (R2010a) software. The images for testing are captured using two Basler-scout cameras (which are used as stereo cameras) mounted on rover at a height of 24.7cm from ground. The specifications (Intrinsic parameters) of the cameras are as follows.

1. Focal length: 0.85 cm
2. Pixel width: 0.465×10^{-3} cm
3. Pixel height: 0.465×10^{-3} cm
4. Resolution:1040 x1392
5. Baseline: 20 cm

The cameras are mounted on rover such that the left camera is rotated 10 degrees anticlockwise along X-direction, rotated 5 degrees clockwise along Y-direction and the right camera is rotated 10 degrees anticlockwise along X-direction, rotated 5 degrees anticlockwise along Y-direction. Two cases are considered here. Firstly, images are captured with pure translation. Second set of images are captured with both rotation and translation. The experimental setup is as shown in Fig. 4.

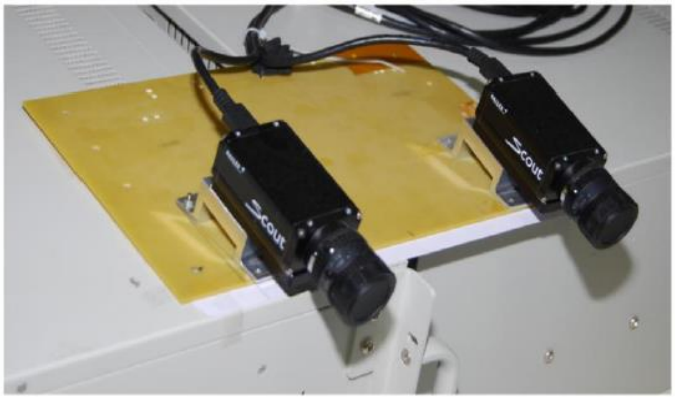

Fig 4: The experimental setup

In order to find out pure translation a set of images are captured at time t=0, t=t+δt. The images captured at time t=0 i.e., at reference images are shown in Fig. 5. The images captured at time $t=t+\delta t_2$ and $t=t+\delta t_2$ are shown in Fig.6 and Fig. 7 respectively.

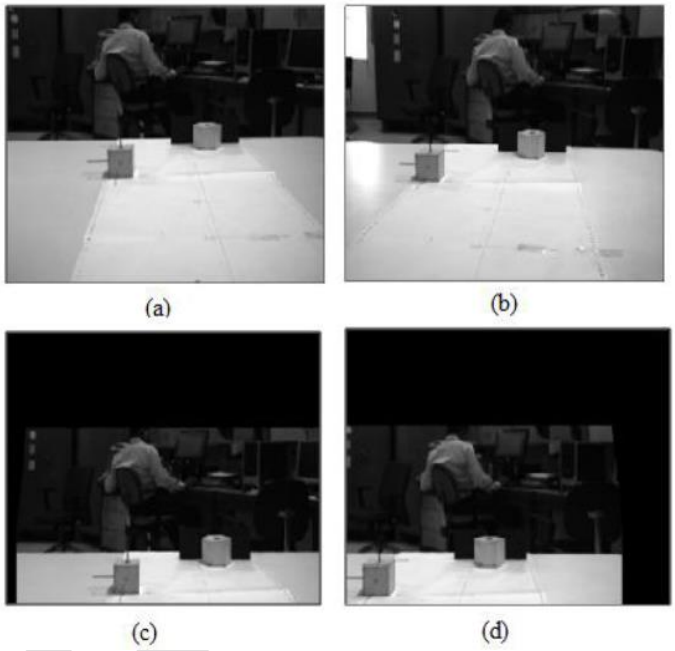

Fig 5: Images for translation at t=0. (a)Reference left image.(b) Reference right image. (c) Rectified reference left image. (d) Rectified reference right image

The images captured at time t=0 are called reference images and the images captured at time t= t+δt are called current images.

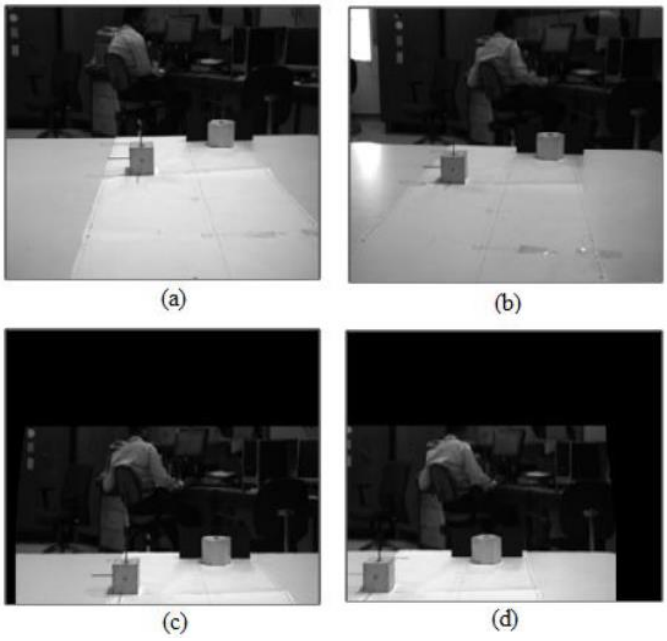

Fig 6: Images for translation at time $t=t+\delta t_2$. (a) Left image.(b) Right image. (c) Rectified left image. (d) Rectified right image.

Table 1 and table 2 gives the comparisons of the estimated and actual translation and depth for 5 feature points for the images hown in Fig. 5 and Fig. 6 respectively. Let Θ, ϕ and ψ are the angles of rotation in degrees along X, Y and Z axis respectively. All the measurements are in cm.

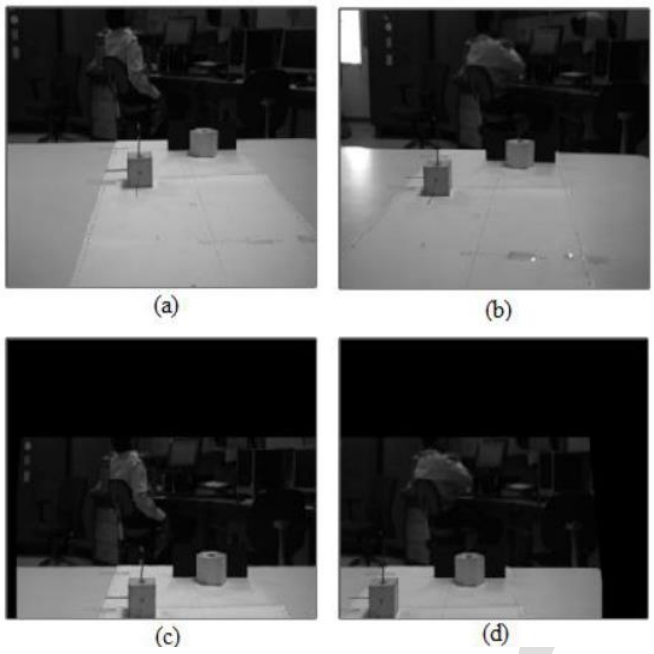

Fig 7: Images for translation at time t=t+ δt_2. (a) Left image.(b) Right image. (c) Rectified left image. (d) Rectified right image.

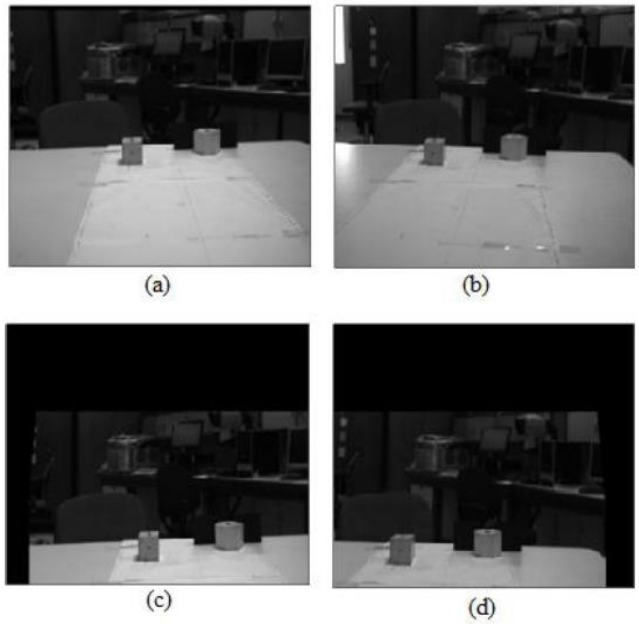

Fig 8: Images for rotation at t=0. (a)Reference left image.(b) Reference right image. (c) Rectified reference left image. (d) Rectified reference right image.

In order to find out rotation with translation a set of images are captured at time t=0, t=t+ δt_1 t=t+ δt_2. The images captured at time t=0 i.e., at reference images are shown in Fig. 8. The images captured at time t=t+ δt_1 and t=t+ δt_2 are shown in Fig. 9 and Fig. 10 respectively. Table 3 and table 4 gives the comparisons of the estimated and actual translation and depth for 5 feature points. Let Θ, ϕ and ψ are the angles of rotation in degrees along X, Y and Z axis respectively. All the measurements are in cm.

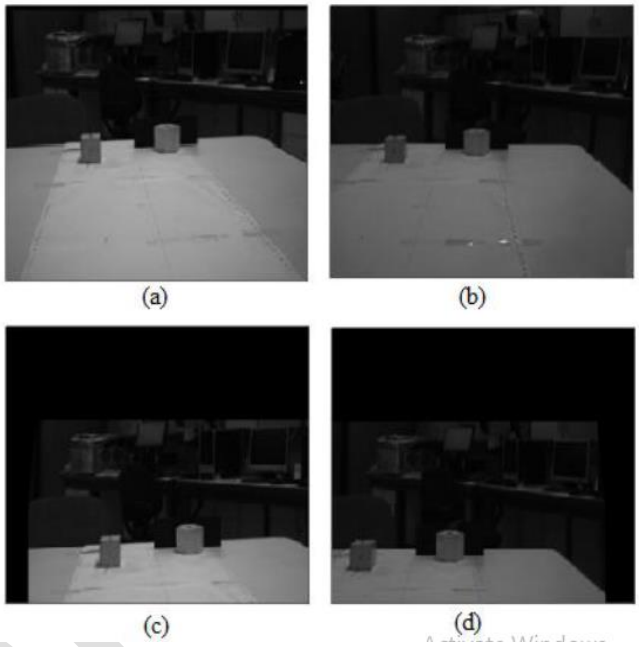

Fig 9: Images for translation at time t=t+ δt_1. (a) Left image.(b) Right image. (c) Rectified left image. (d) Rectified right image.

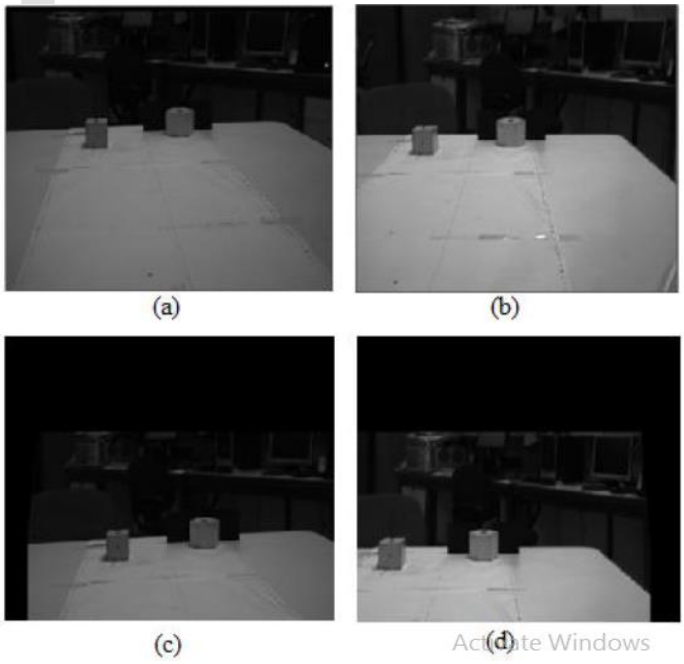

Fig 10: Images for translation at time t=t+ δt_2. (a) Left image.(b) Right image. (c) Rectified left image. (d) Rectified right image.

Table 1: Tabulation of co-ordinates for amount of translation at time $t=t+\delta t_1$ w.r.t t=0

	All measurements are w.r.t Left camera at time t=0						All measurements are w.r.t Left camera at time t=t+1						Translation of Left camera at t=t+1 w.r.t t=0						Rotation of Left camera at t=t+1 w.r.t t=0					
	Measured			Estimated			Measured			Estimated			Measured			Estimated			Measured			Estimated		
FP	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	Θ	ϕ	Ψ	Θ	ϕ	Ψ
1	5	-24.7	90	5.614	-24.310	91.626	0	-24.7	90	0.1504	-24.267	91.626	-5	0	0	-5.435	0.0027	0.2791	0	0	0	0	0	0
2	5	-19.2	90	5.586	-18.5	91.169	0	-19.2	90	0.2	-18.5	91.397	-5	0	0	-5.378	0.0037	0.3454	0	0	0	0	0	0
3	-5	-24.7	120	-4.983	-23.7	119.923	-10	-24.7	120	-9.732	-25.234	119.745	-5	0	0	-5.172	-0.691	-0.334	0	0	0	0	0	0
4	-25	-24.7	120	-25.096	-25.187	119.866	-30	-24.7	120	-29.935	-22.8	118.157	-5	0	0	-5.123	0.052	0.198	0	0	0	0	0	0
5	-10.5	-17	120	-11.012	-17.151	118.5	-15.5	-17	120	-16.025	-17.034	119.234	-5	0	0	-5.1013	0.0183	-0.3193	0	0	0	0	0	0

Table 2: Tabulation of co-ordinates for amount of translation at time $t=t+\delta t_2$ w.r.t t=0

	All measurements are w.r.t Left camera at time t=0						All measurements are w.r.t Left camera at time t=t+1						Translation of Left camera at t=t+1 w.r.t t=0						Rotation of Left camera at t=t+1 w.r.t t=0					
	Measured			Estimated			Measured			Estimated			Measured			Estimated			Measured			Estimated		
FP	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	Θ	ϕ	Ψ	Θ	ϕ	Ψ
1	5	-24.7	90	5.614	-24.3	91.627	0	-24.7	85	0.853	-24.608	85.839	-5	0	-5	-4.27	-0.348	-4.732	0	0	0	0	0	0
2	5	-19.2	90	5.586	-18.5	91.169	0	-19.2	85	0.524	-19.923	86.001	-5	0	-5	-4.703	0.0037	-4.6	0	0	0	0	0	0
3	-5	-24.7	120	-4.98	-23.7	119.92	-5	-24.7	115	-4.781	-23.912	114.32	-5	0	-5	-5.172	-0.691	-5.96	0	0	0	0	0	0
4	-25	-24.7	120	-25.1	-25.2	119.87	-30	-24.7	115	-29.94	-22.8	115.98	-5	0	-5	-5.123	0.052	-4.806	0	0	0	0	0	0
5	10.5	-24.7	90	10.7	-24	90.943	5.5	-24.7	85	5.78	-24.3	86.123	-5	0	-5	-5.101	0.0183	-4.533	0	0	0	0	0	0

Table 3: Tabulation of co-ordinates for amount of rotation at time $t=t+\delta t_1$ w.r.t t=0

	All measurements are w.r.t Left camera at time t=0			All measurements are w.r.t Left camera at time t=t+1			Translation of Left camera at t=t+1 w.r.t t=0			Rotation of Left camera at t=t+1 w.r.t t=0					
FP	X	Y	Z	X	Y	Z	X	Y	Z	Θ	ϕ	Ψ	Θ	ϕ	Ψ
1	-10.0164	-23.67	119.866	0.6711	-23.962	122.6817	1.6925	0.5432	0.9198	0	-5	0	0	-4.505	0
2	-31.238	-24.01	121.234	-20.136	-20.884	118.912	0.8014	0.0037	2.231	0	-5	0	0	-4.921	0
3	-10.0164	-23.67	119.866	0.6711	-23.962	122.6817	2.3052	0.29	2.3602	0	-5	0	0	-4.484	0
4	-17.0323	-18.26	117.9327	-5.6766	-18.192	120.6572	0.2842	0.519	1.6702	0	-5	0	0	-5.386	0
5	-16.879	-23.92	118.312	-5.5738	-24.321	119.866	-5.101	0.0183	-4.533	0	-5	0	0	-5.278	0

Table 4: Tabulation of co-ordinates for amount of rotation at time t=t+ δt_2 w.r.t t=0

	All measurements are w.r.t Left camera at time t=0			All measurements are w.r.t Left camera at time t=t+1			Translation of Left camera at t=t+1 w.r.t t=0			Rotation of Left camera at t=t+1 w.r.t t=0					
FP	X	Y	Z	X	Y	Z	X	Y	Z	Θ	ϕ	Ψ	Θ	ϕ	Ψ
1	-0.4035	-24.553	105.358	10.765	-17.706	107.527	0.9599	-0.5306	1.271	4	-5	0	3.6405	-5.4732	0.346
2	4.7564	-24.47	104.754	15.3936	-17.726	106.586	4.112	0.1904	1.0111	4	-5	0	3.5833	-3.9844	0.251
3	-0.2907	-19.012	106.277	10.2655	-12.035	107.844	-1.0206	-0.5134	1.0993	4	-5	0	4.039	-6.2528	0.3458
4	-17.032	-18.258	117.933	-5.4849	-10.702	122.271	-0.1271	-0.1591	2.3076	4	-5	0	3.7912	-5.4608	0.3454
5	-25.206	-22.603	116.061	-14.027	-15.369	122.682	0.5682	-5.1351	3.0774	4	-5	0	3.591	-5.5269	0.323

IV. CONCLUSION

Stereo vision seems to be a good choice for odometry and it provides an efficient method to estimate the pose and position of autonomous vehicles with good enough accuracy. In this project a set of computer vision techniques are used to find the path traced by the moving vehicle. Histogram equalization or specification, denoising are some of the pre-processing steps. In this project epipolar rectification is done if the stereo images are not rectified, Harris corner detector is used for feature extraction from stereo images, correspondence problem is solved in stereo images using cross-correlation method, triangulation method is used to obtain the 3D co-ordinates of the desired feature point hence finding out the position and QUEST algorithm is used for pose estimation.

The methods which are used in this project are suitable for short distance, high accurate localization. It is based on stereo vision and it has limitations of distance with visibility. Stereo vision systems have a slow working rate, but with the fast evolving technology in the computer vision field, it is assumed that the most of the stereo systems drawbacks can be soon surpassed. This method works efficiently for a range of 3m to 5m. This method also works with the cameras of high resolutions.

REFERENCES

[1]. Zhhiwei Zhu, Taragay Oskiper, Oleg Naroditsky, Supun Samarasekera, Harpreet, Singh Sawhney, Rakesh Kumar, "Stereo-based visual odometry method and system", United States Patent Application Publication, Pub. No US 2008/0144925 a1, Pub. Date Jun.19, 2008.

[2]. "Introductory Techniques for 3D Computer Vision", E.Trucco, A. Verri, Prentice Hall Inc., 7th chapter pp-139-186, 1998.

[3]. "An Introduction to 3D Computer Vision Techniques and Algorithms", Boguslaw Cyganek, J Paul Sibert, Wiley Publication, Pub. No ISBN: 978-0-470-01704-3, Pub. Year 2009.

[4]. Andrea Fusiello, Emanuele Trucco and Alessandro Verri, "A compact algorithm for rectification of stereo pairs", Machine Vision and Applications 12,Springer-Verlag, Page no.16-22, Pub, Year 2000.

[5]. "Image Processing, Analysis, and Machine Vision", Milan Sonka, Vaclav Hlavac, Roger Boyle, 9th chapter, pp-441-502, Pub. No ISBN-0-534-95393-X, Pub. Year 1998.

[6]. Chris Harris and Mike Stephens, "A Combined corner and edge detector", Plessy Research Roke Manor, United Kingdom, The Plessy company, Pub. Year 1988.

[7]. M. D. Shuster and S.D. Oh, "Three-Axis Attitude Determination from Vector Observations", Computer Sciences Corporation, Silver Spring, Md., VOL. 4, NO. 1, Pub. Year 1981.

[8]. F. Landis Markley, "Fast quaternion attitude estimation from two vector measurements", Guidance, Navigation, and Control Systems Engineering Branch, Code 571, NASA's Goddard Space Flight Center, Greenbelt, MD 20771, Pub. Year 1981.