CLUSTERED FEATURES FOR USE IN STEREO VISION SLAM

Deon Joubert[1]
[1] CSIR
Pretoria, South Africa
e-mail1: djoubert@csir.co.za

ABSTRACT

SLAM, or simultaneous localization and mapping, is a key component in the development of truly independent robots. Vision-based SLAM utilising stereo vision is a promising approach to SLAM but it is computationally expensive and difficult to implement. New feature manipulation techniques are proposed which incorporate relational and positional information of the features into the extraction and data association steps.

Keywords: Stereo Vision, Machine Vision, SLAM, Feature Extraction, Data Association.

1 INTRODUCTION

The localization of a mobile robot and positioning of surrounding objects are two problems that must be solved to realize independent robot operation. In a known environment a robot can be localized by determining the relative position of the robot with regards to the known positions of landmarks within the environment. Conversely, if the position of the robot is known, a map can be created of the surroundings by determining landmark positions relative to the robot. For truly independent operation a robot must be able to both create such a map and localize itself in situations where neither the position of the robot nor the position of landmarks are known a priori. Simultaneous localization and mapping (SLAM) is defined as the solving of both problems at the same time.

Robots require sensory input to implement SLAM. Visual sensors are well suited to robotic applications due to their relative inexpensiveness, lightness, low power consumption as well as the high rate of data that cameras produce. The disadvantage of these sensors is the large amount of processing that is required to convert the data into usable information. The extraction of the landmark depth information from images is an especially challenging problem that requires an accurate and efficient solution.

Vision-based SLAM makes use of visual sensors in the implementation of SLAM. A common method employed in vision-based SLAM is to make use of image recognition techniques to find distinguishable features in an image. The three dimensional position of these feature are then determined and then recorded as landmarks within the SLAM map. The feature itself and the processing associated with the feature affect the performance of a SLAM system in both its accuracy and computational efficiency.

2 RELATED WORK

SLAM produces a map which consists of the estimated positions of the landmarks and a covariance matrix which reflects the uncertainty of the landmark positions relative to one another and to the estimated position of the robot. One of the difficulties in the implementation of SLAM is the computational expense of producing this map. Due to the

increasing number of landmarks used and the associated expense of landmark extraction and data association it is found that SLAM systems are often not implementable in real-time, as seen in [1] and [2]. It is therefore important to improve the efficiency of the mapping algorithm as well as to reduce the computation time of any additional processes.

Feature-based visual depth extraction involves the calculation of image depth by relating the difference in position of the same image primitive across several images to the physical configuration or movement of the camera system. Depth is only calculated for a number of points as opposed to the whole of the image and the relationship between features and SLAM landmarks is easily defined. Features that have been used in SLAM implementations include the Harris corner detector ([3], [4] and [5]), scale-invariant feature transform (SIFT) [6], speeded up robust features (SURF) [2] and Lucas-Kanade optical flow [7].

In [8] a data association method is used which was developed specifically for a SLAM system and which utilizes the positional information of the map to improve feature matching. Features are detected using the Shi-Tomasi detector and are stored in a database together with an image patch centred on the corner. The positional estimates of the features are then used to predict where the corner will be visible in subsequent frames, where the robot is in a different position. An image patch of the expected corner position is then transformed using robot positional information to maximize the chances of producing a correct match. This is a very effective method and has been used in several other SLAM implementations, such as [9], [10] and [11].

Point features can be grouped to make use of the geometric relationships between the features to improve feature detection, matching and tracking. SIFT keypoints can be grouped according to objects that are detected in a scene [12]. This approach requires that the SLAM system be provided with a database of SIFT keypoints corresponding to specific objects. This approach is not applicable when the robot has to explore a completely unknown and irregular environment. In [13] the object database instantiation problem is avoided by finding objects in the current environment and adding them to an initially empty database. In [14] groups of SIFT keypoints called fingerprints are used as a way to identify sup-maps in a global map which improves mapping efficiency. In [4] groups of multi-scale Harris corners are matched. Further matching of corners based on the predicted position of corners relative to the matched group is then conducted.

As can be seen, the feature extraction and data association steps in a vision-based SLAM system can benefit from innovative definition and handling of the visual feature. The geometric information of the landmarks can be incorporated into the data association process to improve efficiency and accuracy. Grouping of detected features enables the use of the relationships between the features to improve the detection and matching of features.

3 SYSTEM DESCRIPTION

3.1 Stereo Vision Assembly

The stereo vision assembly consists of two E-54G10HP Power Zoom cameras mounted on a Perspex tray. It allows for various translational and rotational configurations. The

cameras are connected via S-video cables to a Sensoray 2255S USB video capture device which in turn is connected to a computer with a USB cable.

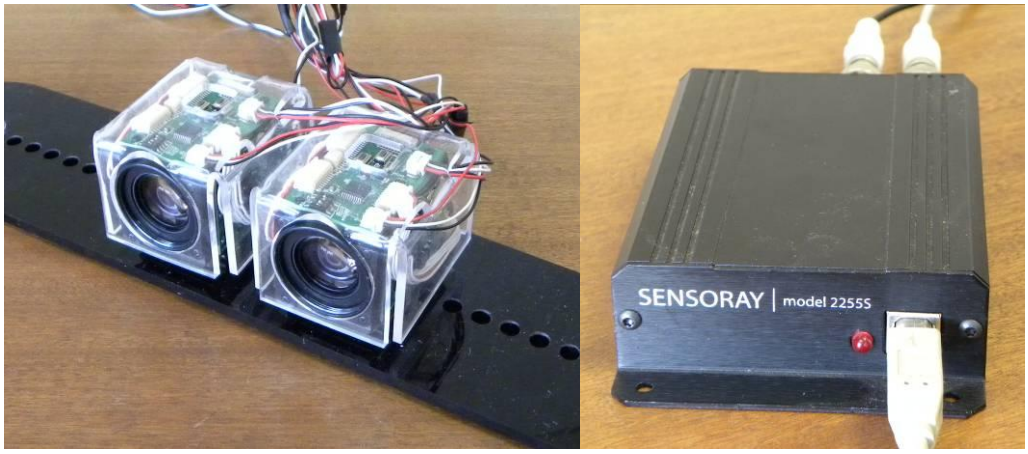

Figure 1. Stereo camera rig and Sensoray 2255S frame grabber

The video capture device, or frame grabber, interfaces with an Ubuntu 8.10 operating system by means of a Video 4 Linux 2 driver. The video stream is accessed and processed by using C based programs and OpenCV. OpenCV is an open source library used for computer vision applications [15].

In order for a system to compute scene depth information from stereo visual input it first needs to determine the geometry of the stereo vision configuration, or the system calibration. Thereafter the video input needs to be adjusted or rectified to compensate for the differences in the two camera orientations. The disparity between the two video inputs can then be calculated by determining the correspondence between the two signals. Knowledge of the geometry of the configuration is then used to transform or reproject the disparity values to depth values.

3.2 Calibration

The geometry of a stereo vision configuration is determined not only by the relative position of the cameras to each other but also by the individual intrinsic parameters of each camera. The intrinsic parameters consist of the focal length and central axis displacement variables. The radial and tangential distortions of the camera influence the rectification process and are also computed during calibration.

The intrinsic parameters relate the spatial position of objects to their position on the image plane. If the object points are known and the image points can be determined, then the intrinsic parameters can be computed. The corners on a grid of black and white squares (a chessboard) are used as object points and corner detection is used to extract the image points. Several views of the chessboard held at different angles must be used for accurate computation of the intrinsic parameters.

After each camera has been calibrated a stereo calibration function determines the rotational and translational matrices that relate the two camera views. This function uses image pairs of a chessboard as captured by the two cameras. The function also further refines the intrinsic camera parameters.

3.3 Rectification

The stereo images are rectified to ease the computation of the image disparities and to compensate for camera distortion of the images. The goal of rectification is to transform the images in such a way that it appears as if the camera were perfectly parallel and horizontally aligned so that the image pixels will be row aligned. The disparity in the scene is then indicated by the difference in the x coordinate of the image screens.

The rectification function is based on Bouguet's algorithm which attempts to minimize the adjustment for each image while maximizing the common viewing area between the images. The function uses the intrinsic parameters for each camera as well as the configuration rotation and translational matrices, as computed during calibration. The algorithm provides the necessary transformation matrices for each camera. It also computes the reprojection matrix to be used during the disparity to depth transformation.

3.4 Correspondence

Stereo correspondence involves determining which points in the two camera images are the same point. If it is known which points are the same, then the disparity between the points can be computed. Feature-based correspondence restricts the matching process to interesting features in each image. Corners are found in each image using the Shi-Tomasi corner detector. A large number of corners are usually found and corners are filtered out according to the quality of the corner as well as the number of corners in an area. Corners are matched between the left and right images using a normalised corner coefficient (NCC) method described in [16]. The disparity between matched corners is then computed.

The result produced by the corner matching algorithm is shown in the following figure.

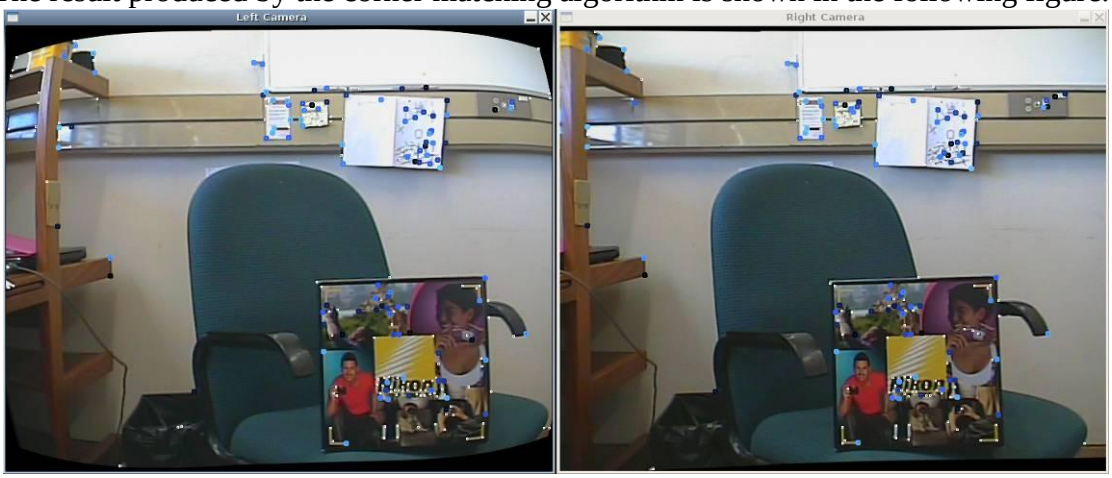

Figure 2. Matching corners are indicated by circles of the same colour.

3.5 Reprojection

The stereo correspondence step provides the coordinates of a number of image pixels in disparity space. These points need to be reprojected to three dimensional space to provide the needed positional information for the SLAM system. The reprojection matrix computed during the rectification step is used to transform the coordinates. This matrix contains the focal length, central axis displacement and distance between the cameras in

an appropriate ordering. The following figure shows an example of depth reprojection. The depth indicated are relative values and not calibrated.

Figure 3. Corner depth is indicated by both colour and rating.

3.6 Landmark Instantiation

Shi-Tomasi corners are detected, matched and projected to a three dimensional space. Corners that are close to one another are clustered together. A hierarchical agglomerative complete-link clustering algorithm [17] is used. Merging of clusters is stopped when the complete-link similarity between clusters reaches a threshold. This threshold is based on experimental observation.

The clusters are now used to denote a number of corners. Clusters are defined by a centre point and a farthest point. The centre point is the mean of the three dimensional positions of the corners in the cluster and the farthest point is the position of the corner situated the farthest from the centre point.

Clusters can be used to speed up feature extraction in the following manner. The coordinates of centre and farthest points are projected to the current disparity space. This disparity space is dependent on the new position of the robot and the new orientation of the cameras. The distance between the image centre and farthest points can then be used to determine the size of the region of interest in which the corners should be searched for. Instead of trying to match the previously detected corners to newly detected corners in the whole of the image, corners are now matched in a smaller search space.

Figure 4. Corner clusters indicated by large circles.

The region of interest is valid only for the left camera image. The region of interest for the right camera needs to be determined. Corners are clustered according to spatial coordinates. The larger clusters are expected to be found on surfaces perpendicular to the line of sight and should have similar disparity values. Therefore the adjustment that must be made to the region of interest for the right camera image can be computed using the projected centre point disparity value.

The clusters can also be used to determine the areas in a view where features are sparse. As each cluster defines a region where corners are detected, the regions of the image where regions have not been declared can then be declared as unknown regions. Feature extraction applied to these areas would lead to a more comprehensive map and better localization. It is also possible to lower the limits of the detection and matching filters so that more corners with less quality can be detected in areas where detection is difficult. In such a case a measure of the quality of the corners in the cluster should be recorded for use in the SLAM system.

4 CONCLUSION

Feature management techniques have been developed to improve the accuracy and performance of the feature extraction and data association processes of a SLAM system. These techniques are based on feature clustering and the utilisation of the recorded feature positional information. The system described in this paper is still a work in progress but it is believed that, if refined, the techniques can greatly improve the performance of a robot employing vision-based SLAM. The future goal is to incorporate multiple features into a single landmark which would improve the fundamental operation of a SLAM system.

5 RECOMMENDATIONS

The Shi-Tomasi corners can be replaced with a visual feature that is invariant to scale changes, such as SIFT or SURF. Furthermore, the adjustment of the region of interest for the right camera relies on a questionable assumption and it should be further developed.

6 REFERENCES

[1] Bryson, M., Sukkarieh, S., Building a robust implementation of bearing-only inertial SLAM for a UAV, Journal of Field Robotics, 2007.

[2] Mahon, I., Williams, S. B., Pizarro, O., Johnson-Robertson, M., Efficient view-based SLAM using visual loop closures, IEEE Transactions on Robotics, 2008.
[3] Thornqvist, D., Schon, T. B., Karlsson, R., Gustafsson, F., Particle filter SLAM with high dimensional vehicle model, Journal of Intelligent and Robotic Systems, 2009.
[4] Lemaire, T., Berger, C., Jung, I., Lacroix, S., Vision-Based SLAM: Stereo and Monocular Approaches, International Journal of Computer Vision, Vol. 74, No. 3, United States of America, 2007.
[5] Artieda, J., Sebastian, J. M., Campoy, P., Correa, J. F., Mondragon, I. F., Martinez, C., Olivares, M., Visual 3-D SLAM from UAVs, Journal of Intelligent and Robotic Systems, 2009.
[6] Zhou, W., Miro, J. M., Dissanayake, G., Information-efficient 3-D visual SLAM from unstructured domains, IEEE Transaction on Robotics, 2008.
[7] Elmogy, M., Zhang, J., Robust real-time landmark recognition for humanoid robot navigation, Proceedings of the 2008 IEEE International Conference on Robotics and Biomimetics, Bangkok, Thailand, 2009.
[8] Davison, A. J., Reid, I. D., Molton, N. D., Stasse, O., MonoSLAM: Real-time single camera SLAM, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 6, 2007.
[9] Paz, L. M., Pinies, P., Tardos, J. D., Neira, J., Large-scale 6-DOF SLAM with stereo-in-hand, IEEE Transactions on Robotics, Vol. 24, No. 5, 2008.
[10] Gemeiner, P., Ponweiser, W., Einramhof, P., Vincze, M., Real-time SLAM with a high-speed CMOS camera, 14th International Conference on Image Analysis and Processing, 2007.
[11] Sola, J., Monin, A., Devy, M., Vidal-Calleja, T., Fusing monocular information in multicamera SLAM, IEEE Transactions on robotics, Vol. 24, No. 5, 2008.
[12] Choi, J., Lee, K., Ahn, S., Choi, M., Chung, W. K., A practical solution to SLAM and navigation in home environment, SICE-ICASE International Joint Conference, 2006.
[13] Lee, Y., Song, J., Visual SLAM in indoor environments using autonomous detection and registration of objects, IEEE International Conference on Multisensor fusion and Integration for Intelligent Systems, 2008.
[14] Schleicher, D., Bergasa, L. M., Barea, R., Lopez, E., Ocana, M., Nuevo, J., Fernaindez, P., Real-time stereo visual SLAM in large-scale environments based on SIFT fingerprints, IEEE International Symposium on Intelligent Signal Processing WISP, 2007.
[15] Bradski, G., Kaehler, A., Learning OpenCV: Computer vision with the OpenCV library, O'Reilly Media Inc., Sebastopol, 2008. 08.
[16] Agrawal, M., Konolige, K., Bolles, R. C., Localization and mapping for autonomous navigation in outdoor terrain: a stereo vision approach, IEEE Workshop on Applications of Computer Vision, 2007.
[17] Jain, A. K., Murty, M. N., Flyn, P. J., Data clustering: a review, Association for Computing Machinery Surveys, 1999.