International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-4, Issue-3, March 2018 Pages 46-50

# Visual SLAM and Vision-Based Localization for Humanoid Robots

**Titus Iulian Ciocoiu, Florin Dumitru Moldoveanu**

***Abstract* -In this article will be presented the problem of real-time localization for humanoid robots. For this purpose we using a single camera as the only sensor. In order to obtain fully autonomous robots an accurate localization of the robot in the world is much more than desirable. Moreover, if we can obtain an accurate localization in real-time, than we can use the remaining computational resources to perform other important humanoid robotic tasks such as planning [1], 3D object modeling [2] or visual perception [3].**

***Index Terms*–Accurate localization in real-time, computational resources, humanoid robots, map of the environment, small laser range sensors.**

## I. INTRODUCTION

An accurate and fast localization of the robot will be a great benefit for many humanoid robotics applications. In order to obtain a robust localization, there are different alternatives of which we can choose: one of them is to estimate simultaneously the localization of the robot and the map of the environment, yielding the well-known Simultaneous Localization and Mapping (SLAM) problem in the robotics community [4]. Another possible option is to dedicate more computational resources in the reconstruction of a persistent map, and using then this map for long-term localization or navigation purposes.

In order to do this, we can take advantage of the prior map of the robot's environment learning different parameters ranging from visibility prediction [5].

In the particular case of humanoid robots, it is very important that the sensors to be light-weight and small. Humanoids should be stable under all possible motions, and heavy sensors can compromise this stability. Furthermore, not all sensors are suitable for humanoid robots. For example not all laser scanners can be mounted on humanoid platforms, especially the heavy ones such as for example the SICK LMS-221. Only small laser range sensors (e.g. Hokuyo URG-04LX) are suitable for humanoid robotics applications [6].

The main problem of these small laser range sensors is the limited distance range (up to 4 m for the Hokuyo URG-04LX). For all of these reasons, cameras are an appealing sensor for humanoid robots: they are light-weight and cheaper than laser scanners and stereo cameras can also provide higher distance ranges (depending 77 78 Visual SLAM and Vision-Based Localization for Humanoid Robots on the stereo rig baseline). More than this, most of the advanced commercial humanoids platforms are already equipped with vision systems. Anyway, there have been only limited attempts at vision-based localization for humanoid robots.

The purpose of this article is to show that is possible to obtain a real-time robust localization of a humanoid robot, with an accuracy of the order of cm just using a single camera and a single CPU. To do this, prior to localization we have to compute an accurate 3D map of the environment using the stereo visual SLAM algorithm described in Chapter 2. To build an accurate 3D map we use stereo vision mainly for two reasons: to be able to measure in a direct way the scale of each detected point and to obtain dense depth information, which is a well-studied problem for stereo vision [7]. In this way we can solve the main drawback of monocular SLAM approaches, i.e. recovering the scale of a map due to observability problems in recovering 3D information from 2D projections.

When we have the 3D map of the environment, we can perform the monocular vision-based localization using the 3D map as a prior. For the localization experiments the dense depth map generation process have to be avoided, which can be a high time consuming operation in certain occasions;then we can perform a robust and efficient real-time localization just using a single camera.

In order to satisfy all of these demands, firstly has to be built a 3D map of the environment using stereo visual SLAM techniques based on Bundle Adjustment (BA). Being inspired by the recent works in nonlinear SLAM, we will use a stereo visual SLAM algorithm combining local BA and global BA to obtain an accurate 3D maps with respect to a global coordinate frame. These maps can be used later for monocular vision based localization or navigation. In this way, 3D points and camera poses are refined simultaneously through the sequence by means of local BA, and when is detected a loop closure, the residual error in the reconstruction can be corrected by means of global BA adding the loop closure constraints.Once obtained, the 3D map of the environment canbe used for different robotics applications such as localization, planning or navigation. Vision-based localization in a large map of 3D points is a challenging problem. One of the most computationally expensive steps in vision-based localization is the data association between alarge map of 3D points and 2D features perceived by the camera. Then, matching candidates are usually validated by geometric constraints

**Titus Iulian Ciocoiu,** Automatics and Applied Informatics, Transilvania University, Brasov,Romania
**Florin Dumitru Moldoveanu,** Automatics and Applied Informatics,TransilvaniaUniversity,Brasov,Romania

using a RANdomSAmple Consensus(RANSAC) framework [8]. Therefore, we have to find a smart strategy to sample the large database of 3D points and perform an efficient data association between the 3D map points and perceived 2D features by the camera. Given a prior map of 3D points and perceived 2D features in the image, our problem to solve is the estimation of the camera pose (with known intrinsic parameters) with respect to a world coordinate frame. Basically, this problem is known in the literature as the Perspective-n-Point (PnP) problem [9]-[11].

For solving efficiently the PnP problem, will be used the visibility prediction algorithm described in [5].Visibility prediction exploits all the geometric relationships between camera poses and 3D map points in the prior 3D reconstruction. Then, during vision-based localization experiments the data association and robot localization will be speeded-up tremendously by predicting only the most highly visible 3D points given a prior on the camera pose. In this way, the PnP problem can be solved in a more efficient and faster way, reducing considerably the number of outliers in the set of 3D-2D matching correspondences.

## II. Monocular Vision-Based Localization

Once obtained the 3D map of the environment (by using the stereo visual SLAM algorithm), the following step is to use that map for common humanoid robot tasks such as navigation or planning, while providing at the same time an accurate localization of the robot with respect to a global coordinate frame. To do this, the obtaining of a real-time and robust vision-based localization is mandatory. Given a prior map of 3D points and perceived 2D features in the image, the problem to be solved is the estimation of the camera pose with respect to the world coordinate frame. Basically, the problem we have to solve now is known as the Perspective-n-Point (PnP) problem.

The PnP problem estimating the pose of a calibrated camera based on 2D measurements and known 3D scene, is a thoroughly studied problem in computer vision [9], [10]. This is generally a challenging problem, even with a perfect set of known 3D-2D correspondences. Although there are some globally optimal solutions [11] that employ Second Order Cone Programs (SOCP), the main drawback of the current globally optimal solutions to the PnP problem is the computational burden of these methods. This makes difficult to integrate these algorithms for real-time applications such as the ones we are interested with humanoid robots.

The main contribution of our work for solving the PnP problem efficiently, is the use of the output of the visibility prediction algorithm (given a prior on the camera pose) to predict only the most highly visible 3D points, reducing considerably the number of outliers in the set of correspondences. In this way, we can make the data association between 3D map pointsand 2D features easier, thus speeding up the pose estimation problem. In figure 3.1is described an overall overview of our vision-based localization approach with visibility prediction.

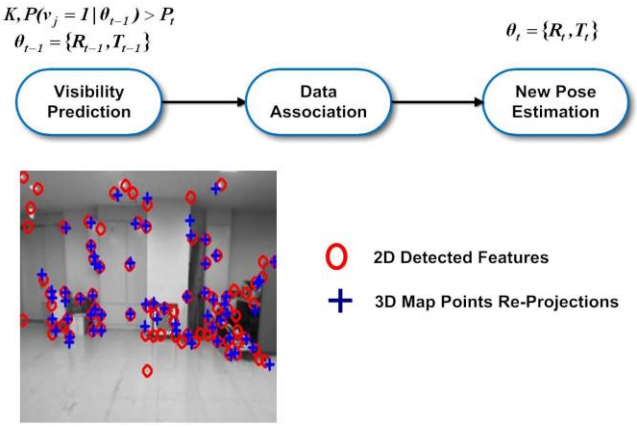



Figure 1: The input for the visibility prediction algorithm is the latest camera pose θt−1, the number of KNNs (K) and a probability threshold Pt. Only the highly visible 3D map points are re-projected onto the image plane of the left camera, and a set of putative matches between 2D detected features and map elements is formed. Then, the PnP problem is solved yielding the localization of the robot with respect to a world coordinate frame $\theta_t$at time t.Best viewed in color.

To clarify, the overall vision-based localization algorithm works through the following steps:

1. While the robot is moving, the camera acquires a new image from which a set of image features $Z_t$= {$z_t$,1. . . $z_t$,n} are detected by a feature detector of choice. A feature descriptor is computed then for each of the detected features. Notice, that even any kind of feature detector and descriptor may be used, it is necessary that both detector and descriptor are the same and have the same settings as in the map computation process described in Chapter 2.
2. By using the visibility prediction algorithm, a promising subset of highly visible 3D map points is chosen then and re-projected onto the image plane based on the estimated previous camera pose θt−1 and known camera parameters.
3. Afterwards, a set of putative matches Ct are formed where the i-th putative match Ct,iis a pair {$z_t$,k,$x_j$} which comprises of a detected feature $z_k$and a map element $x_j$. A putative match is created when the Euclidean distance between the appearance descriptors of a detected feature and a re-projected map element is lower than a certain threshold.
4. The pose estimation problem is finally solved by minimizing the following cost error function, given the set of putative matches Ct:

$$\arg\min_{R,T} \sum_{i=1}^{m} ||z_i - K(R * x_i + t)||_2 \quad (1)$$

where zi= (uL, vL) is the 2D image location of a feature in the left camera, xi represents the coordinates of a 3D point in the global coordinate frame, K is the left camera calibration matrix, and R and t are respectively the rotation and the translation of the left camera with respect to the global coordinate frame. The PnP problem is formulated asa non-linear least squares procedure using the LM algorithm implementation describedin [12]. The set of putative

matches may contain outliers, therefore RANSACis used in order to obtain a robust model free of outliers.

## III. INITIALIZATION AND RE-LOCALIZATION

During the initialization, the robot can be located in any particular area of the map. In order to do this we need to find a prior camera pose to initialize the vision-based localization algorithm. For this purpose, we have to compute the appearance descriptors of the detected 2D features in the new image and match this set of descriptors against the set of descriptors from the list of stored key frames from the prior 3D reconstruction. In the matching process between the two frames, it is performed a RANSAC procedure forcing epipolar geometry constraints. The camera pose is recovered from the stored key frame that obtains the highest inliers ratio score. If this inliers ratio is lower than a certain threshold, then the localization algorithm do not have to be initialized until the robot moves into a known area, yielding a high inliers ratio. At this point, we are confident about the camera pose prior and initialize the localization process with the camera pose parameters of the stored key frame with the highest score.

At this point it may happen eventually that the robot gets lost due to bad localization estimates or that the new camera pose is rejected due to a small number of inliers in the PnP problem. In those cases, must to be performed a fast re-localization by checking the set of appearance descriptors of the robot's new image against only the stored set of descriptors of the key frames that are located in a certain distance area of confidence around the last accepted camera pose estimate.

## IV. RESULTS AND DISCUSSIONS

Hereinafter will be described one of the experiments conducted on the HRP-2 humanoidrobot. We created for it two different datasets of common humanoid robotics laboratory environments. Here will be presented the first dataset which is called Tsukuba, and it was done at the Joint RoboticsLaboratory, CNRS-AIST, Tsukuba, Japan. This dataset comprises of different sequencesfor the evaluation of the monocular vision-based localization algorithm under the assumptionthat a prior 3D map is known. In particular, in this dataset we have different robottrajectories (square, straight) and challenging situations for the localization such as robotkidnapping, people moving in front of the robot and changes in lighting conditions. Forthis dataset, we performed experiments with an image resolution of 320×240 and a framerate of 15 frames per second. The main motivation of using that image resolution is thatin this dataset we focused more on achieving real-time localization results while at thesame time obtaining robust pose estimates.

In the Tsukuba dataset, the experiments were performed considering an image resolution of 320×240 and a frame rate of 15 frames per second. For the visibility prediction algorithm we consider the following input parameters of the algorithm: K = 10 and Pt >0.20. We chose a threshold value of 2 pixels in the RANSAC process, for determining when a putative match is predicted as an inlier or outlier in the PnP problem.

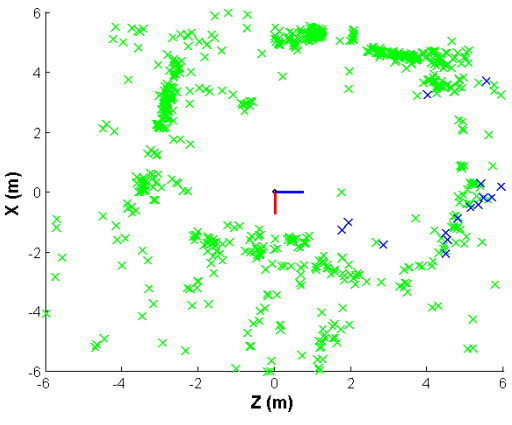
Square 2 m Size Sequence.In this sequence, the robot performed a 2 m size square in a typical humanoid robotics laboratory. This sequence was designed for capturing different camera viewpoints both intranslation and orientation. Firstly, we built a 3D map of the environment by using the stereo visual SLAM algorithm described in Chapter 2 in a similar sequence and performedvisibility learning. The resulting 3D map comprises of 935 points and 75 key frames. At the start of the sequence, we placed the robot at the origin of the map, and then by using the pattern generator, the robot performed a square of 2 m size. We measured manually the final position of the robot, and this position was (X = 0.14, Y = 0.00,Z= −0.02) in meters. Due to the existing drift between the planned trajectory and the real one, the robot was not able to close the loop itself. Then, we validate our vision-based localization algorithm with visibility prediction under a similar square sequence.

Fig.3.1 depicts the initial and final position of the robot, and the performed trajectory.

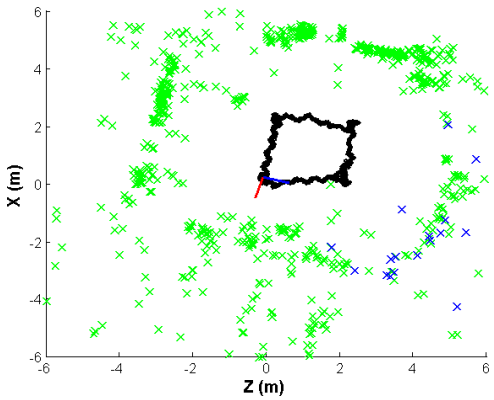
Table 3.1 shows the obtained localization results using visibility prediction forthis square sequence. According to the results we can see that the localization accuracyis very good, about the order of cm. The differences with respect to the real trajectoryfor the final position are very small 9 cm, in the X coordinate and about 7 cm in the Zcoordinate. While the robot was walking the pattern generator fixed the Y coordinatealways to the same value. Therefore, in the PnP problem we add this constraint to speedup the process, although our algorithm can deal with 6DoF.

Table 3.1: Square 2 m size monocular vision-based localization results (Tsukuba dataset).

| Camera Pose Element | Start Position | Final Position |
|---|---|---|
| X (m) | 0.0000 | 0.2320 |
| Y (m) | 0.0000 | 0.0000 |
| Z (m) | 0.0000 | -0.0092 |
| $q_0$ | 1.0000 | 0.9905 |
| $q_x$ | 0.0000 | 0.0034 |
| $q_y$ | 0.0000 | 0.1375 |
| $q_z$ | 0.0000 | 0.0050 |




(a)




(b)

Fig.2 : Square 2 m size localization results Tsukuba dataset. (a) and (b) depict the initialand final position of the robot and the performed trajectory in the sequence. In these two imagesthe robot trajectory is depicted in black, the visible 3D points are depicted in blue and the rest of 3D points in green.Best viewed in color.

## V. CONCLUSION

In this paper we have presented a vision-based localization algorithm that works in real-time (even faster than 30 Hz) and provides localization accuracy about the order of cm. Firstly was built a 3D map of the environment by using stereo visual SLAM techniques, and perform visibility learning over the prior 3D reconstruction. For a faster visionbased localization we use then visibility prediction techniques for solving the PnP problem and obtaining the location of the robot with respect to a global coordinate frame. The accuracy of our localization algorithm was measured by comparing the estimated trajectory of the robot with respect to ground truth data obtained by a highly accurate motion capture system. In addition, we are interested in improving the capabilities of our vision-based localization algorithm towards the goal of life-long localization and mapping. We are also interested in combining visibility prediction with the Bayesian Surprise and landmark detection framework [13]. We can make a model in a probabilistic way when the robot discovers a new surprising area and then adding this area into the whole reconstruction. Indeed, we also think that Bayesian surprise can be also useful for detecting a new object in the prior 3D reconstruction, and once the robot detects this new object, the robot can start a 3D reconstruction of the object using the localization information as a prior.

We have mainly put our focus in real-time vision-based localization. We think that the accuracy in localization can be increased if we fuse the information from our vision-based localization with the odometry information of the robot. The image resolution and length of the descriptors can be increased also, but the price to pay is higher computational demands, that may prevent the algorithm from real-timeperformance. In any way, the main limitation is the higher computational demands of these kind of invariant feature detectors. In the next future, we are interested in using our approach in related vision-based humanoid robotics problems such as control [14], autonomous 3D objectmodeling [2] or footstep planning [1]. We like to think that our real-time vision based localization can improve considerably some previous humanoidrobotics applications where vision-based localization was not exploited in all its capabilities.

## ACKNOWLEDGMENT

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), ID137070 financed from the European Social Fund and by the Romanian Government.

## REFERENCES

[1] Perrin, N., Stasse, O., Lamiraux, F., and Yoshida, E. (2010). Approximation of feasibility tests for reactive walk on HRP-2. In IEEE Intl. Conf. on Robotics and Automation (ICRA), pages 4243–4248, Anchorage, AK.

[2] Foissote, T., Stasse, O., Wieber, P., Escande, A., and Kheddar, A. (2010). Autonomous 3D object modeling by a humanoid using an optimization-driven next-best-view formulation. Intl. J. of Humanoid Robotics.

[3] Bohg, J., Holst, C., Huebner, K., Ralph, M., Rasolzadeh, B., Song, D., and Kragic, D. (2009). Towards grasp-oriented visual perception for humanoid robots. Intl. J. of Humanoid Robotics, 6(3):387–434.

[4] Durrant-White, H. and Bailey, T. (2006). Simultaneous localization and mapping SLAM:part 1. IEEE Robotics and Automation Magazine, 13(3):99–110.

[5] Alcantarilla, P., Ni, K., Bergasa, L., and Dellaert, F. (2011). Visibility learning in largescale urban environment. In IEEE Intl. Conf. on Robotics and Automation (ICRA), Shanghai, China.

[6] Kagami, S., Nishiwaki, K., Kuffner, J., Thompson, S., Chestnutt, J., Stilman, M., and Michel, P. (2005). Humanoid HRP2-DHRC for autonomous and interactive behavior. In Proc. of the Intl. Symp. of Robotics Research (ISRR), pages 103–117.

[7] Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Intl. J. of Computer Vision, 47:7–42.

[8] Bolles, R. and Fischler, M. (1981). A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In Intl. Joint Conf. on AI (IJCAI), pages 637–643, Vancouver, Canada.

[9] Lu, C., Hager, G., and Mjolsness, E. (2000). Fast and globally convergent pose estimation from video images. IEEE Trans. Pattern Anal. Machine Intell., 22(6):610–622.

[10] Ansar, A. and Danilidis, K. (2003). Linear pose estimation from points or lines. IEEE Trans. Pattern Anal. Machine Intell., 25(4):1–12.

[11] Schweighofer and Pinz, 2008.

[12] Lourakis, M. (2004). levmar: Levenberg-Marquardt nonlinear least squares algorithms in C/C++ [web page]http://www.ics-forth.gr/~lourakis/levmar/.

[13] Ranganathan, A. and Dellaert, F. (2009). Bayesian surprise and landmark detection. In IEEE Intl. Conf. on Robotics and Automation (ICRA), Kobe, Japan.

[14] Blösch, M., Weiss, S., Scaramuzza, D., and Siegwart, R. (2010). Vision based MAV navigation in unknown and unstructured environments. In IEEE Intl. Conf. on Robotics and Automation (ICRA), Anchorage, AK, USA.

**Titus Iulian Ciocoiu**is a third yeardoctoral student active in the SystemEngineering research field at the“Transilvania” University of Brasov,Romania, The working title of his thesis is“Security in digital electronic systems”. His research is focused on digital imageprocessing and active vision systems. Heholds a diploma in Automatic and AppliedInformatics and a master in AdvancedAutomatic Systems and InformationTechnology both from the “Transilvania”University of Brasov.

**Florin Moldoveanu** received the B. Sc., and Ph. D. degrees in electrical engineering fromTransilvania University of Brasov, Romania,in 1975 and 1998, respectively. He iscurrently Professor as part of the Departmentof Automation and Information Technology,Faculty of Electrical Engineering andComputer Science, Transilvania University of Brasov. His main research interests focus ondigital circuits, discret event systems, sliding mode control, digital image processing techniques.