Optical Sensing for Robot Perception and Localization

○Yutaka YAMAMOTO, Paolo PIRJANIAN, Joe BROWN, Mario MUNICH, Enrico DiBERNARDO,
Luis GONCALVES, Jim OSTROWSKI, Niklas KARLSSON
Evolution Robotics, Inc.
130 W. Union Street,
Pasadena, CA 91103
yutaka.yamamoto@evolution.com

Abstract: Optical sensing, e.g., computer vision, provides a very compelling approach to solving a number of technological challenges for developing affordable, useful, and reliable robotic products. We describe key advancements in the field consisting of three core technologies for Visual Pattern Recognition (ViPR), Visual Simultaneous Localization and Mapping (vSLAM), and a low-cost solution for localization using optical beacons (NorthStar).

ViPR is an algorithm for **Vi**sual **P**attern **R**ecognition based on scale invariant features (SIFT features) which provides a robust and computationally effective solution to fundamental vision problems including the correspondence problem; object recognition; structure; and pose estimation. vSLAM is an algorithm for Visual Simultaneous Localization and Mapping using one camera sensor in conjunction with dead-reckoning information, e.g., odometry. vSLAM provides a cost-effective solution to localization and mapping for cluttered environments and is reliable to dynamic changes in the environment. Finally, NorthStar uses IR projections onto a surface to estimate the robot's pose based on triangulation. We give examples of concept prototypes as well as commercial products such as Sony's Aibo, which have incorporated these technologies in order to improve product utility and value.

Introduction

A key attribute of autonomous intelligent robots is reliable perception in real world environments. Once a robot is able to reliably perceive its environment using its sensors it will be able to perform intelligent actions to accomplish important and useful tasks such as navigation, human-robot interaction, and manipulation of the environment. These are tasks that are fundamental to many applications of mobile robots ranging from robotic vacuum cleaning, hospital delivery robots, elder care robots, etc. In addition to reliability of perception, for widespread acceptance of such applications it is necessary that the underlying technologies provide an affordable solution, i.e., the component technologies have to be cost-effective. Our approach to providing affordable and cost-effective solution to perception is to use optical sensing, e.g., computer vision, to solve problems of perception and localization. Due to the massive use of cameras in applications such as digital photography and camera phones, the price of the imaging sensor has decreased significantly and made them very attractive for consumer products. In addition, cameras can be very versatile sensors that can be used to solve a number of key problems for robotics and other automated and intelligent systems. They can be used for navigation, face recognition, gesture recognition, object recognition, and much more. They key challenges to taking advantage of this powerful but yet inexpensive sensor is to come up with algorithms that can reliably and effective extract the information necessary for solving these problems. Typically, image processing algorithms can be fragile to changes in illumination, clutter in the environment, and other visual artifacts. Furthermore, image processing algorithms can be computationally intensive and hence require powerful and hence expensive computing resources.

In this paper, we describe optical solutions to the problems of object recognition and localization using reliable and cost-effective technologies. We describe a Visual Pattern Recognition technology which is reliable to changes in lighting, partial occlusions, rotation, and affine distortions. Further, we describe vSLAM, a solution for localization and mapping using one camera. vSLAM provides a solution to localization that is two orders of magnitude less expensive compared to SLAM techniques using laser range finders. Finally, we describe another approach to localization which uses optical beacons in the environment to estimate the pose of the robot using a very inexpensive optical sensor.

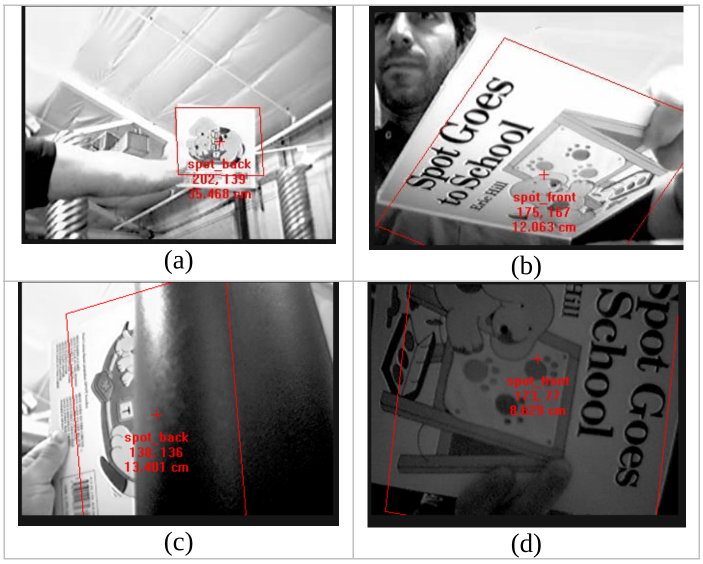
Visual Patter Recognition

Our Visual Pattern Recognition algorithm, ViPR, is a vision-based module that can be trained to recognize objects using a single, low-cost camera. The main strengths of the object recognition module lie in its robustness in providing reliable recognition in realistic environments where, for example, lighting can change dramatically. ViPR consists of representing an object as a set of localized visual templates (also called features) extracted from one or more images of such object at multiple levels of resolution. The key elements of the algorithm consist in (1) the particular choice of features to be used, called SIFT(Scale Invariant Feature Transform) features, which are highly localized, invariant to changes in scale and rotation and partially invariant to changes in illumination and viewpoint, (2) the very efficient way of organizing and searching through a database of hundreds of thousand of SIFT features, and (3) the robust matching technique used to match images to models via geometric alignment and voting. The matching technique is particularly robust to occlusion. ViPR has a recognition rate of 80-95% in uncontrolled settings and up to 95% rate in more controlled settings with false positive rates well below 1%. In addition to identification of targets, ViPR estimates the pose of the camera with respect to the object of interest. This capability is used by the Sony AIBO to successfully localize and navigate towards the charging station.

Figure 1 illustrates the performance of the resulting algorithm. The ER Vision Object Recognition software can provide a recognition rate of 80-95% in uncontrolled settings and up to 95% rate in more controlled settings with false positive rates well below 1%.

The recognition speed is a logarithmic function in the number of objects in the database. The object library can store hundreds or even thousands of objects without a significant increase in computational requirements. The recognition frame rate is proportional to CPU power and image resolution. For example, the recognition rate is 5 frames per second (fps) with an image resolution of 320x240 on an 850MHZ Pentium III processor and 3 fps at 80x66 on a 100MHz MIPS-based 32-bit processor. Reducing the image resolution decreases the image quality and, ultimately, the recognition rate. However, performance degrades gracefully with decreasing image quality. Each object model requires about 40KB of memory.

ViPR was released as part of Sony AIBO® vision system in year 2003 and aids AIBO to return to its charging station as well are for recognizing flash cards for human-robot interaction.



(a) (b) (c) (d)

Figure 1. (a) ViPR is able to recognize the object at various distances. (b) it can deal with rotations and affine transformations. (c) it can handle a large amount of occlusions. (d) it can handle a large change in illumination.

Visual Simultaneous Localization and Mapping

Localization is the process of a robot estimating its own position with respect to its environment. In pure localization, however, the robot must be provided an accurate map. This is clearly a constraining element because 1) the robot cannot adapt to changes in the environment, e.g., after reorganizing furniture or remodeling, and 2) manually building a map is a labor-intensive task, which may require specific expertise and hence add to installation cost. In order to reach adaptability and a higher degree of autonomy the robot must build a 'map' using its sensors, and then to use this map to find out where it is. This is known as *Simultaneous Localization and Mapping*, or SLAM for short.

One approach to moving away from costly range sensors is to use cameras for localization and mapping. Recent results reported on vision-based localization [1], however, rely on a known map and are constrained to semi-structured environments. *Visual SLAM* (vSLAM) provides a first-of-a-kind, cost-efficient solution to localization and mapping that is a major breakthrough for robotics in general and consumer robotics in particular. By using a low-cost camera and optical encoders as the only sensors this solution can reduce the cost of localization by 2 to 3 orders of magnitude compared to existing range-based techniques. The breakthrough that vSLAM provides consist not only of a major cost improvement but also of a significant improvement in performance. Since it relies on visual information rather than range measurements it can be used in cluttered environments such as furnished homes and office buildings, where range-based SLAM techniques tend to fail. Furthermore, our vSLAM algorithm is highly robust to dynamic changes in the environment caused by lighting changes, moving objects and/or people. By combining reliable performance with low-cost requirements vSLAM is an important and enabling, breakthrough technology for consumer robotics.

This novel approach to localization and mapping builds a visual map, which consists of a set of unique landmarks created by the robot along its path. A landmark consists of an image of the scene along with the robot's position to indicate the position estimate of the landmark. The *i*th landmark is described by a tuple $L_i = <I_i, S_i>$, where I_i is the image of the landmark and S_i is the position estimate of the landmark. A map, m, consists of a set of k landmarks, $m = \{L_1, L_2, \ldots, L_k\}$.

In an unknown environment, the robot automatically generates new landmarks, which are maintained in a database. Initially, the position estimates of the landmarks are based on wheel odometry alone. As the robot revisits mapped areas, it uses known landmarks to estimate an improved robot pose, which is then used to improve the landmark position estimates. By using the landmarks for robot pose estimation vSLAM can bound the error introduced by odometry drift.

vSLAM continuously compares the current image, seen from the robot, with the images of the database to find matching landmarks. Such a match is used to update the pose of the robot according to the relative position of the matching landmark. Furthermore, it updates the position of the matched landmark using the robot's current pose estimate. By continuously updating the position of landmarks based on new data, vSLAM incrementally improves the map by calculating more accurate estimates for the position of the landmarks. An improved map results in more accurate robot pose estimates. Better pose estimates contribute to better estimates for the landmark positions and so on. So vSLAM continuously improves the map and the robot's pose estimates based on newly acquired data. Our experimental results from various sites show an accuracy of about 10cm in *x* and *y* position estimates and an accuracy of about 5 degrees in absolute heading estimates.

When the robot does not see or detect any known landmarks then it takes two specific actions. First it updates the robot's pose estimate using odometery measurements relative to the last position estimate. Wheel odometry can be quite accurate for short traversals but can introduce unbound errors for longer travel distances. When the odometry error grows large then vSLAM updates the map with a new landmark. In this way, vSLAM gradually builds a denser map but not denser than necessary since landmarks are only created when no others are found. As a result, vSLAM is very adaptive to changes in its environment. I.e., if the environment changes so much that the robot no longer recognizes previous landmarks then it automatically updates the map with new ones. Outdated landmarks that are no longer recognized can easily be deleted from the map by simply keeping track of if they were seen/matched when expected.

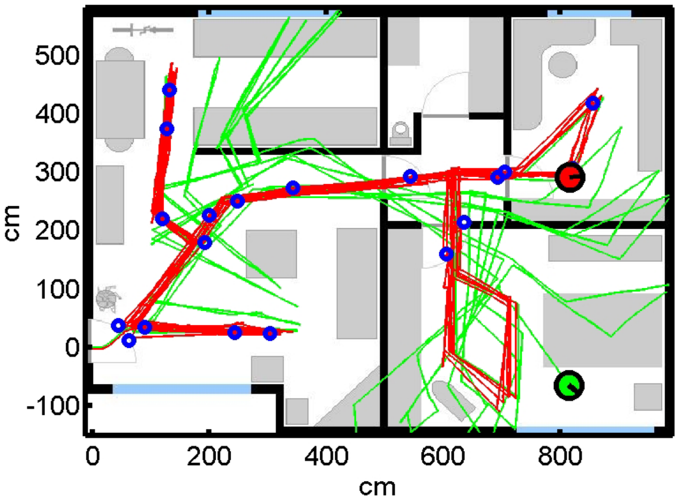


Figure 2. The figure shows the path and the visual map generated by vSLAM in a 2 bedroom apartment. The red path is the one generated by vSLAM. The green path is the one generated by wheel odometry. The blue circles are the position of the landmarks generated by vSLAM. They layout of the apartment was manually superimposed on the visual map as a point of reference. This layout was not generated by vSLAM.

Figure 2 shows the results of vSLAM from a run in a two bedroom apartment. The robot was driven around along the dashed reference path (this path is unknown to the SLAM algorithm). The vSLAM algorithm builds a map consisting of landmarks marked with blue circles in the figure. The corrected robot path, which uses a combination of visual landmarks and odometry, provides a robust and accurate position determination for the robot as seen by the red path in the figure. Clearly, the green path (odometry only) is incorrect since, according to this path, the robot seems to be traversing through walls and furniture. The red path (the vSLAM corrected path), on the other hand, is consistently following the reference path.

Beacon-Based Localization

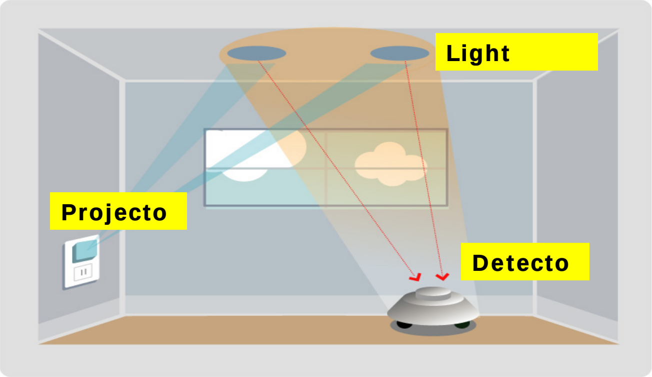


Figure 3. Conceptual view of NorthStar. The projector creates two unique IR spots in the ceiling which the detector on the robot uses for triangulation to recover the pose of the robot.

Northstar is an optical solution to localization which consists of a beacon which projects unique patterns on a surface, preferably the ceiling of a room, and an optical sensor such as a camera to detect the projected patterns based on their surface reflections. Based on detection of the position of two spots/patterns it is possible to use triangulation to estimate the position of the camera within an environment. Northstar is an attractive solution for consumer robotics products because: a) it can be realized at very low-cost; b) it is minimally intrusive, and c) it is not limited by line of sight availability between the emitter and the detector. NorthStar provides effective localization for most indoor environment with position accuracy of about 10cm in x, y and 5 degrees in heading in the effective field of view of the sensor which is about 4m^2.

A major disadvantage of beacon-based solutions to localization is that they require installation and perhaps maintenance of beacons in the environment. Reducing the intrusiveness of such solutions is very important. A key advantage of NorthStar over other beacon-based localization solutions is that its beacons can be easily installed in an environment by just plugging the projector into a wall outlet. Another key advantage of NorthStar over other beacon-based solutions is that it is not limited by line of sight. Typically, optical beacon-based solutions require line of sight visibility between the emitter-detector pair which limits their range of operation in most environments due to occlusions caused by furniture and other objects in the environment. Since NorthStar localization is based on detecting surface reflection of the patterns rather than direct detection of the emitter we can minimize the line-of-sight limitation. The key idea is to minimize or reduce the line-of-sight limitation by projecting the light sources onto a surface that is visible from a relatively large portion of the environment. For example, in an indoor environment, it can be advantageous to project the emitted light from the beacon onto the ceiling. In many indoor environments, the ceiling of a room is visible from most locations with the room.

Position estimation

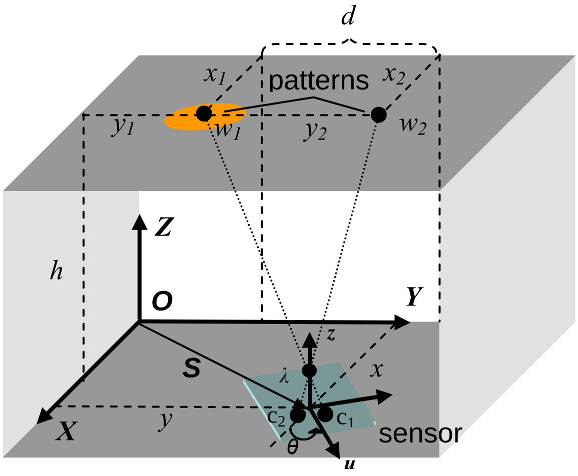


Figure 4. The geometrical model used to relate the NorthStar projections on the ceiling to the NorthStar sensor.

The NorthStar sensor can be a CCD camera or a CMOS camera and one or more projectors each of which generates one or more patterns on the projection surface, e.g., the ceiling. The projector can for example consist of a laser pointer device retrofitted with a lens that creates a unique pattern. E.g., one pattern could have the shape of a square and another pattern could have the shape of a circle. Each camera will generate grayscale or color images. A signal processing unit will process the camera images to extract the unique patterns and estimate their position in image coordinates. The position of each pattern can be defined as the centroid of the pattern. Let w_j define the position of the jth pattern and let c_i define the position of the jth pattern in the ith camera's image coordinates (see figure 4). Then the relationship between the jth pattern and its projection on the ith camera will be defined by equation (1).

$$c_i = \boldsymbol{P}\boldsymbol{R}_{\boldsymbol{\theta}} (w_i - S) \qquad (1)$$

where S is the position of the camera, $\boldsymbol{P}$ is defined as the perspective transformation from a world point (X, Y, Z) to the corresponding image point (u, v) assuming that the camera coordinate system is aligned with the world coordinate system. For a pinhole camera model $\boldsymbol{P}$ is defined as $\lambda/(\lambda\text{-}Z)$, where λ is the focal length of the camera. $\boldsymbol{R}_{\boldsymbol{\theta}}$ is defined as a rotation matrix.

Then based on equation (1) we can solve for the position of the camera, S, we get the following equation:

$$S = w_i - \boldsymbol{P}^{-1}\boldsymbol{R}_{\boldsymbol{\theta}}^{-1} c_i \qquad (2)$$

Once the image position of the patterns are extracted in image coordinates then the signal processing unit can estimate the position of the camera in a similar manner as described by equations (2). Also note that the projectors can on-off modulate the projection of the patterns to reduce the effects of ambient light. The modulation frequency can be used to associate a unique ID to each pattern, however, note that the ID of each pattern can be encoded in the pattern itself. E.g., the shape of the pattern can define a unique ID for each pattern if distinct shapes are used for each pattern. For example, the system will be able to distinguish between the square pattern from the circle pattern and associate different IDs to each.

Conclusions

Inexpensive and reliable perception is a key component in making robots autonomous and intelligenet for real world applications. Optical sensors, such as cameras, are powerful in solving robot perception because of the versatility of the sensors and the low-cost. In this paper, we describe three core technologies that provide very powerful capabilites for robots: object recognition and localization. These technologies have been commercialized and implemented in advanced robotic products such as the Sony AIBO entertainmetn robot. With the low cost implementation of vSLAM and NorthStar we believe that the furture robotic vacuums do not need to resort to random coverage strategies.

References:

1. Wolf, J., Burgard, W. and Burkhardt, H. *"Robust Vision-based Localization for Mobile Robots Using an Image Retrieval System Based on Invariant Features",* Proceedings of the 2002 IEEE Int. Conf. on Robotics and Automation, Washington, DC, USA, May, 2002. P.359-363.