Selforganizing Visual Perception for Mobile Robot Navigation

Georg von Wichert
Control Systems Theory & Robotics Department
Technical University Darmstadt
Landgraf-Georg-Str. 4, D-64283 Darmstadt, Germany
gvw@rt.e-technik.th-darmstadt.de

Abstract

Visual navigation for mobile robots is almost always performed on the basis of complex CAD models which have to be given to the system in advance. There are some approaches that map visual information more or less directly to robot actions in order to implement some basic capabilities like road following or docking. Complex navigation tasks, e.g. in buildings, however, are usually solved using three dimensional models of the environment or the objects within it. This paper presents a method for selforganizing the visual perception of a mobile robot in order to adapt it to the surroundings without the need to define and model the relevant aspects of the environment. The system uses a selforganization process to transform the continuous flow of images into a limited number of discrete perceptions which can be used for navigation purposes. Results obtained in an indoor office environment are presented. Nevertheless, the approach is suited also for outdoor navigation.

1. Introduction

Mobile robots are the subject of large scientific interest since a long time. In most cases they have been the reference testbed for autonomous systems research. The technologies so far developed to reach the goal of mobility and autonomy are currently used for new applications rising in the field of service robotics. These new applications with an increasing demand of robot–environment interaction rely on powerful perception capabilities. Ultrasonic sensors or one dimensional laser range finders, which have been widely used for pure transportation and navigation tasks, will in most cases not be sufficient. Therefore systems using 2D and 3D cameras have become the focus of many research projects. However, most vision guided robots depend on models a-priory known to the system, either of the environment itself or objects within it.

Service robots will, in many cases, need the ability to work in environments which are unknown at the design time of the robot and can therefore not be modeled in advance. This is also true for the environmental features and their sensor images that could be used by the system, for example as landmarks. Modeling will be restricted to very global and syntactical information (i.e., that the environment consists of different rooms linked by doors).

This paper presents a method for selforganizing the visual perception of a mobile robot to adapt it to the surroundings without the need to define and model the relevant aspects of the environment. The system is able to transform the continuous flow of images by means of a selforganization process into a limited number of discrete perceptions which can be used for navigation purposes.

The paper is organized into 5 sections. Section 2 shortly summarizes previous work on the subject before section 3 gives details of our approach. Section 4

Figure 1. The mobile robot ALEF
then describes the mobile robot system ALEF (fig. 1) and presents some experimental results obtained in a normal and unprepared office environment. Finally the paper is summarized in section 5 and an outlook on future work is given.

2. Related Work

In the past sensor based navigation systems typically relied on ultrasonic sensors or laser scanners providing one dimensional distance profiles. The major advantage of this type of sensors results from their ability to directly provide the distance information needed for collision avoidance. In addition, various methods constructing maps and environment models needed for long term as well as for short term (reactive) planning have been developed.

The only but important drawback of these sensors is a consequence of the trivial statement that, using the normal sensor arrangement, only vertical structures, i.e. mainly the shape of the free space surrounding the robot, can be recognized. The real world service applications envisaged in most of the current research projects, however, demand more detailed sensor information to enable the system to really interact with its environment. This can only be derived from high resolution image sensors like 2D and 3D cameras.

Due to tremendous amount of work currently being done in this field it is not possible to discuss all related work completely. Therefore we shortly present three different groups of systems that can be identified in literature. We do not take into consideration the systems navigating within the conventional traffic on normal roads but stay in the classic robotics field:

- A large number of systems navigate with conventional distance sensors and use vision to find and identify objects to be manipulated.
- There are several systems which directly couple the sensor output to motor control in a supervised learning process. The goal is to learn basic skills like road following[9] or docking[2]. Typically non complex scenes are presented to these systems, allowing simplifying assumptions concerning the image processing itself.
- Vision sensors produce a huge amount of data which have to be examined. To reduce the massive data flow, most systems really navigating using cameras, predefine relevant low level (e.g edges, planes) and high level (e.g. objects, doors, etc.) features [6][1] to be included in the environment model. These basic features have to be chosen and modeled in advance. In addition a sensor processing system has to be established to guarantee proper identification of the modeled features in the sensor data. Some of the systems aim at constructing models for unknown environments once the landmarks have been modeled [12].

There exists some work that tries to reduce the modeling effort by interactive construction of geometric environment models [8]. Nevertheless the problem in principle still remains.

![Figure 2. Topological map building by vector quantisation of the (ultra-sonic) sensor data space by means of a selforganizing feature map. This work shows how these ideas can be extended to use video images.](image)

The work of Kurz [7] has shown that basic tasks like the navigation in unknown environments do not necessarily require a model based interpretation of the sensor data. His system can, using a ring of ultrasonic sensors, construct topological maps (fig. 2) suited for navigation based on an internal representation of the environment. The system itself decides which environmental features can be extracted from the sensor data and therefore are used for navigation purposes.

In the following sections we will show that this can also be done for video images. We emphasize the fact that all processes run unsupervised and autonomously. At no level designer defined models or structural assumptions concerning the environment are used.

3. Selforganization of Perceptual Capabilties

The final goal of every image or scene analysis is to find a set of features which characterize the scene under
3.1. Unsupervised Segmentation

In step one the image is segmented in an unsupervised process to stabilize the sensor image against small changes, e.g., in camera position. As depicted in figure 4 this is done by first computing a set of local (to every pixel) features. For grey level images texture features (e.g., the local texture energy features by Laws [4] or a similar method from [10]) are suited. For color images alternatively features derived from the color information could be used. These features are then arranged into multi-channel feature images which are segmented by quantization of the feature vectors for each pixel. This is done by means of a hierarchically structured set of Kohonen's feature maps.

Figure 5 shows the segmentation result in comparison with the input image. Areas of homogeneous local texture features (in this image Laws texture energy features) are plotted using the same color depending upon the position of the winner neuron in the feature map.

This so-called class image is stable against small changes of camera position or lighting conditions. Form, size and (texture) class of the segmented areas are almost independent from the pose of the area within the image. As a result, the compact image description which is computed in the next step, will

---

1 We use a multi-resolution approach.
smoothly change when moving the robot smoothly. This is crucial for topological navigation based on these image descriptions.

3.2. Image Description

After having stabilized the incoming sensor information, a compact description of the resulting "class images" has to be generated. This image description shall afterwards be used to compare images and group them into a finite set of discrete perceptions. It is obvious to choose the form of a vector for this description. By this a characteristic vector will be assigned to every image. A second vector quantization will group the images using these scene feature vectors.

In general, there is of course an unlimited number of ways to construct such vectors. We have used several possible methods ([11]), but the best results were obtained using two dimensional geometric moments (up to third order in the experiments) of the distributed areas belonging to the same texture class as extracted by the previous step. This method is often used for shape recognition ([5]). When computing the moments, the foveal region is emphasized by a sigmoid weight function in order to smooth the resulting image description when moving the robot. The scene feature vector is then composed from the scalar moment values for all classes.

The grouping of various images into discrete perceptions is done by a quantization of the scene feature vectors. In our experiments this quantization is performed by means of a Growing Neural Gas Network (GNG Network) ([3]), which, in contrast to the classical self-organizing feature map by Kohonen, is able to choose the number of neurons dynamically during training. In addition, GNG networks learn the topology of the input data without the need to specify the dimensionality of the data in advance; topological artifacts are thus avoided by definition (see figure 6, from [3]). As a result of this, the system is able to choose the number of different perceptions (i.e., neurons) autonomously during the training process. Of course the number of neurons can be limited. Some changes to the algorithm by Fritzke have been implemented to adapt the network to the needs of robotics applications. They are quite similar to those proposed in [13].

4. Experiments

This section presents some results obtained with our approach in an unprepared office environment. Section 4.1 describes our mobile robot ALEF which was used in the experiments, Section 4.2 the results.

4.1. Experimental Setup

Figure 1 shows our mobile robot ALEF. It is based on a RWI-B12 platform and equipped with 24 ultrasonic sensors plus a CCD camera. A 486-based on
board computer is responsible for basic tasks like ultrasonic sensor based collision avoidance, odometric navigation and radio communication with an off board control computer. Via an analog radio link the video images are transferred to the same off board computer. Thus it is possible to perform all high level tasks off board.

The experiments were carried out in a normal office environment where the robot traveled for some time along the path plotted in figure 7. The small arrows indicate the robot position and orientation at the moment of image acquisition. 700 images were recorded during the experiment. The path led the robot through three rooms and a corridor. Some typical images are given in figures 8 through 9.

### 4.2. Results

After training, each of the neurons in the GNG network represents a group of optically similar images. The sum of them implements an internal environment representation which groups parts of the continuous flow of images to a set of discrete perceptions. The level of detail depends upon the number of neurons allowed. Figure 8 shows that the images are distinguished very roughly when only few neurons are used. Images of two classes are given along with an interpretation.

When we allow the system to use more neurons (see fig. 9 where 16 classes have been distinguished), single classes specialize on certain image types (e.g. the shelf) which are also recognized when traveling along different paths. This can also be noticed in figure 10 which shows an enlarged portion of the path where the robot is traveling in both directions of our corridor. Arrows again indicate the robots position and orientation at the moment of image acquisition, while colors code the class number of the image, i.e. show the discrete perceptions. For the left result 15 classes (neurons in the GNG net) were used for classification and images in both directions are perceived as the same, while using 50 classes (on the right) leads to different perceptions depending upon the driving direction. Note, that the number of subsequent images grouped together does not change. This can be interpreted as a hint for the adequacy of the robot's internal representation of environment and sensing capabilities.

Because of its unsupervised nature the classification process does not always lead to these transparent and "easy-to-interpret" results\(^2\). Interpretations like those

\(^2\)This is not necessary, however, as this is the robots internal
Figure 8. Rough distinction between 'Free Space' and 'Narrow Passage' using just few (3) classes.

Figure 9. A specialist emerging when using many classes. The specialized shelf class recognizes the shelf on different passes.

Figure 10. Enlarged portion of the path. Arrows indicate the robot's position and orientation at the moment of image acquisition. Colors code the class number of the image.

given in figure 8 are thus generally not allowed. Nevertheless, the results are plausible to a human observer in most cases.

The computation time needed to perform the overall processing of the images is 3.06 sec/image on one SPARC-Processor (SPARC 20, 60 MHz).

5. Conclusions and Future Work

This paper presents a new approach to vision-based navigation. The mobile robot ALEF is able to self-organize its visual perception for navigation purposes. No models of the environment or objects within it are used. The system is completely unstructured and makes no restricting assumptions concerning the environment and its structure. Figure 11 demonstrates that the proposed unsupervised image segmentation mechanism works also for outdoor images that are completely unstructured. This segmentation result can be improved further by selecting other local feature filters.

The methods discussed above enable the robot to extract a finite set of discrete perceptions from the more or less continuous flow of images. It somehow learns a set of terms for the description of the surrounding world. This is the basic prerequisite for any further map building and planning.

The terms generated up to now only depend on the input data. They are completely "meaningless". The meaning of any sensor information can only be derived when the system's options to interact with the environment are also taken into account.

In addition, the terms refer to an internal representation and are not very transparent to a human user of representation of the world.
such a system (see section 4.2). To communicate with the robot it will be necessary to define a common level for robot and user. This will not be done by simply labeling all image classes with user defined attributes, but by giving the robot a syntactical description of the environment, e.g., "the world consists of rooms and corridors which are linked by doors". Knowing this syntax the robot has to apply it to the current operating environment with only minimal help (e.g., answers to certain questions) by the user.

Future work will concentrate on these two fields, the improvement of the self-organizing perception by including the systems options to interact and the definition of a system-user cooperation level.

References


Figure 11. Results from the office environment can in general be transferred also to other worlds.