Speech and Image Recognition for Intelligent Robot Control with Self Generating Will

Peter Nauth
Fachhochschule Frankfurt a.M. – University of Applied Sciences
Nibelungenplatz 1, 60318 Frankfurt, Germany
pnauth@fb2.fh-frankfurt.de

Abstract – Intelligent Robots need to recognize a goal to be met and their environment in order to achieve the goal. By means of speech recognition and image processing an intelligent humanoid robot has been developed which can understand spoken commands, sense the environment in order to identify objects of interest and execute the instruction a user has spoken to the robot. Additionally, an algorithm to equip the robot with a self generating will is proposed.

Keywords - Goal Achievement; Goal Understanding; Humanoid Robots; Image Processing; Robot Control; Speech Recognition; Self Generating Will; Sensor Fusion.

I. Introduction

Robots of the next generation such as assistive robots or rescue robots must be able to solve complex tasks which up to now only human beings can handle. They will act autonomously in a natural environment and will communicate in a natural way with those people they are supposed to support.

Algorithms for robot control and navigation have been developed by several research groups [5]. This paper focuses on the understanding of goals, strategies to achieve these goals, learning methods, on sensing and navigating in a natural environment as well as on handling objects with respect to intelligent humanoid robots. In order to cope with situations where the robots are confronted with conflicting requirements, drives and experiences, a self-generating will is of advantage as opposed to a rigid state-action architecture.

Another focus is on applying the algorithms to small robots. The advantages of small sized robots over other systems [5] are reasonable deployment costs and scalability.

II. System Architecture

An autonomous robot needs to know the goal to be accomplished, situation awareness and the ability to plan and perform actions depending on a situation. This requires the following functions [8]:

- Sensing by means of multiple sensors in order to acquire all necessary data about the environment. It includes getting to know the goal to be met, e.g. by understanding a spoken instruction.
- Fusion of the data acquired from intelligent sensors in order to assess the situation
- Planning how to achieve the goal and
- Execution of the necessary steps by controlling the robot motors

In order to sense the environment, the robot is equipped with the following sensors (Fig. 1):

- Speech Recognition Sensor
- Proximity Sensor
- Smart Camera.

The signal processing for speech recognition is implemented on the DSP-based module Voice Direct 364 (Sensory Inc.). The image processing algorithms run on the smart colour camera POB-Eye (POB Technologies) with an embedded 32-Bit controller ARM7TDMI. It is mounted on the neck of the humanoid robot and can be turned left and right as well as up and down.

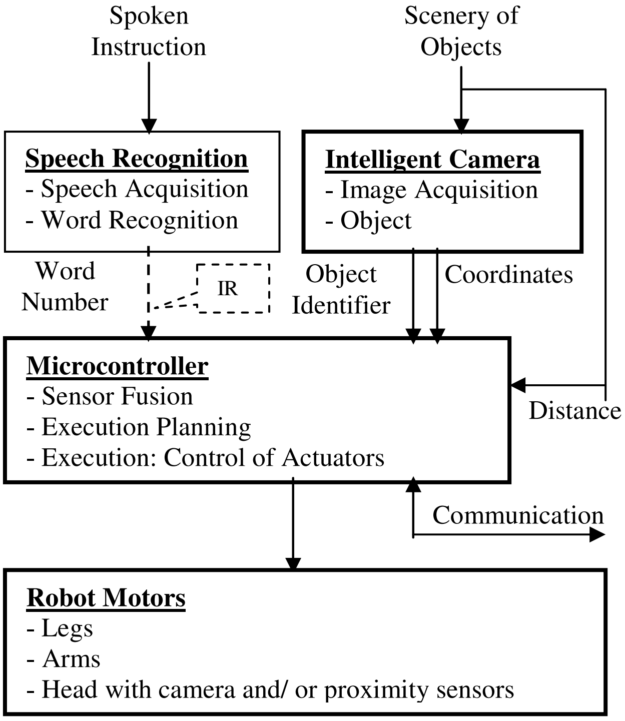

Figure 1. System Architecture: Thick lines indicate parts mounted at the robot

In order to cope with new situations, the speech sensor can learn new words and the vision sensor can learn shapes and colours of new objects [1].

Two different proximity sensors (laser beam triangulation method) mounted on the moveable head measure the distance to the object in the near range from 4 to 24 cm and the far range from 15 to 80 cm. Additionally, they acquire information about the object's shapes.

III. Speech Recognition

The intelligent speech recognition enables the robot to understand spoken instructions. These are either single words or a sequence of words which are spoken without breaks in between. After data acquisition, the algorithm divides the signal into segments and calculates the frequency spectra out of each segment. Next, frequency parameters are calculated and classified by means of a neural network.

As for the training phase, the user speaks a word and repeats it. If the frequency parameters from the first and the repeated word are similar, the word is accepted. The weighting factors of the classifier are being adapted to the word's frequency parameters and assigned to a word number. Then the training can be continued with the next word. By saying "Say a word" the speech sensor asks the user to speak his instruction during the recognition phase. If the classifier detects a high similarity to one of the previously learned words, it sets a respective pin to High. An additional controller SAB 80535 turns the spoken word into a word number by monitoring the state of the pins and transmitting a bit sequence via an infrared (IR) LED to the robot. The sequence corresponds to the pin number set to high and therefore to the word number recognized by the speech module. This enables the user to command the robot remotely.

The robot controller receives the bit sequence via an infrared detector and decodes the word number. For each word number is related to an instruction, the robots now knows its goal, i.e. which object to search for.

IV. Object Recognition

For the recognition of the demanded object and for obstacle avoidance during the search phase the intelligent camera and the proximity sensors are used.

The algorithms we have developed for the smart camera converts the acquired RGB – image into the HSL – space, segments the image [3] by means of an adaptive threshold algorithm (hue histogram) and extracts the form factor F

$$F = U^2 / A$$

from the area A and the circumference U as well as the mean hue – value H out of each object detected. By means of a box classifier each object is assigned to an object identifier which represents the class (Fig. 2). Given the extracted parameters, objects and obstacles can be differentiated regarding shape and colour.

Additionally, the coordinates of each object are calculated. The object identifier and the respective coordinates of all objects found are transmitted to the robot microcontroller.

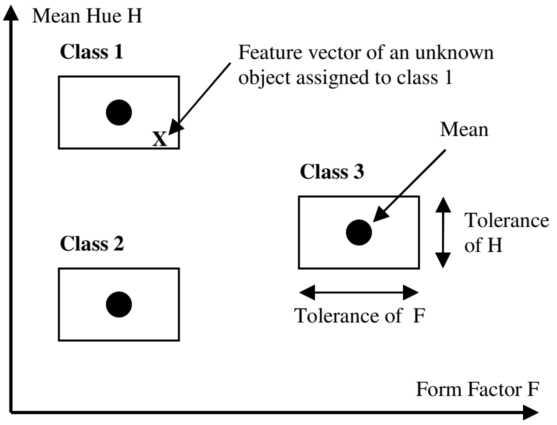

Figure 2. Box Classifier: Mean values und tolerances result from the learning phase

New objects can be learned by a supervised learning algorithm: Typical examples of each object class are shown to the camera and the learning algorithm assigns the mean values of each parameter to the class these objects belong to. The tolerances which define the size of the classification box of each class equal 1.5 times the standard deviation calculated during the teach-in procedure.

Proximity sensors supplement the camera information by sensing the distance between the robot and the object. Additionally, proximity sensors can be used to provide shape parameters for object differentiation themselves, especially for obstacle detection. Given that the proximity sensors scan the environment in two dimensions, the distance z (α,ß) is a function of the horizontal angle α and the vertical angle ß. Appropriate parameters can be selected by modelling the distance function for typical object shapes. E.g., the distance function z (α,ß) of a round object with the radius r positioned in a lateral distance d can be modelled as

$$z(\alpha,\beta) = \frac{(d+r)\cos\alpha - \sqrt{(d+r)^2(\cos^2\alpha - 1) + r^2}}{\cos\beta}$$

whereas the distance function z (α,ß) of a wall is

$$z(\alpha,\beta) = \frac{d}{\cos\alpha\cos\beta}$$

Stairs can be detected and differentiated from objects with a smooth shape such as bottles and walls if unsteady parts exist in the distance function z (α,ß) acquired when turning the head vertically.

V. Sensor Fusion, Planning and Motion Control

By fusing the auditive, visual and proximity data, the robot knows all objects within its reach and their position as well as the goal it is advised to attain.

The fusion algorithm used is hierarchical and works as follows:

1. Auditive and visual data are fused by matching the word number (speech sensor) with one of the object identifiers (camera) by means of a table and generate one of the following hypothesis:
    - Object not found
    - Object found
    - Wall or Stairs
2. The robot moves towards the object or obstacle. Next, the hypothesis generated by the visual sensor is verified by the data acquired from the proximity sensor. If the class derived from the distance function z (α,ß) equals the hypothesis it is regarded as true. Otherwise it is rejected.

In order to execute the appropriate steps for goal achievement the robot has to plan the next actions necessary to meet the goal. First, the robot assesses its state $s_j(\underline{x})$ which is a function of the sensory input $\underline{x}$. Next, the action

$$a_i\left(s_j(\underline{x}), g, a_{i-k}, \underline{x}\right)$$

is derived depending on the actual state $s_j(\underline{x})$, the goal g to be met, previous actions a_{i-k} and the sensory input $\underline{x}$. After having executed the action a_i the robot reaches the next state $s_{j+1}(\underline{x})$.

Depending on the state which has been reached after an action, the robot develops different plans. If it had performed the action "sensor fusion" the plan would be:

- If the demanded object has been identified, the robot approaches it, grabs it and delivers.
- If no or the wrong object has been spotted or in case of conflicting results, the robot repeats the search.
- If an obstacle has been detected the robot develops an approach to overcome it, i.e. it climbs stairs or avoids colliding with walls.

VI. Application Example

One typical example of the robot's performance is to search for objects and bring it to the user (Fig. 3).

If the user says "Water Bottle", the robot understands its task and searches for the bottle. After detection, it approaches the bottle, grabs it and brings it to the user. Even stairs can be climbed by a coordinated arm and leg movement and a stabilization of the robot with a tilt sensor.

In a scenario with a wall, stairs and 3 different bottles, the right bottle was found in 18 cases. The experiment has been repeated 20 times.

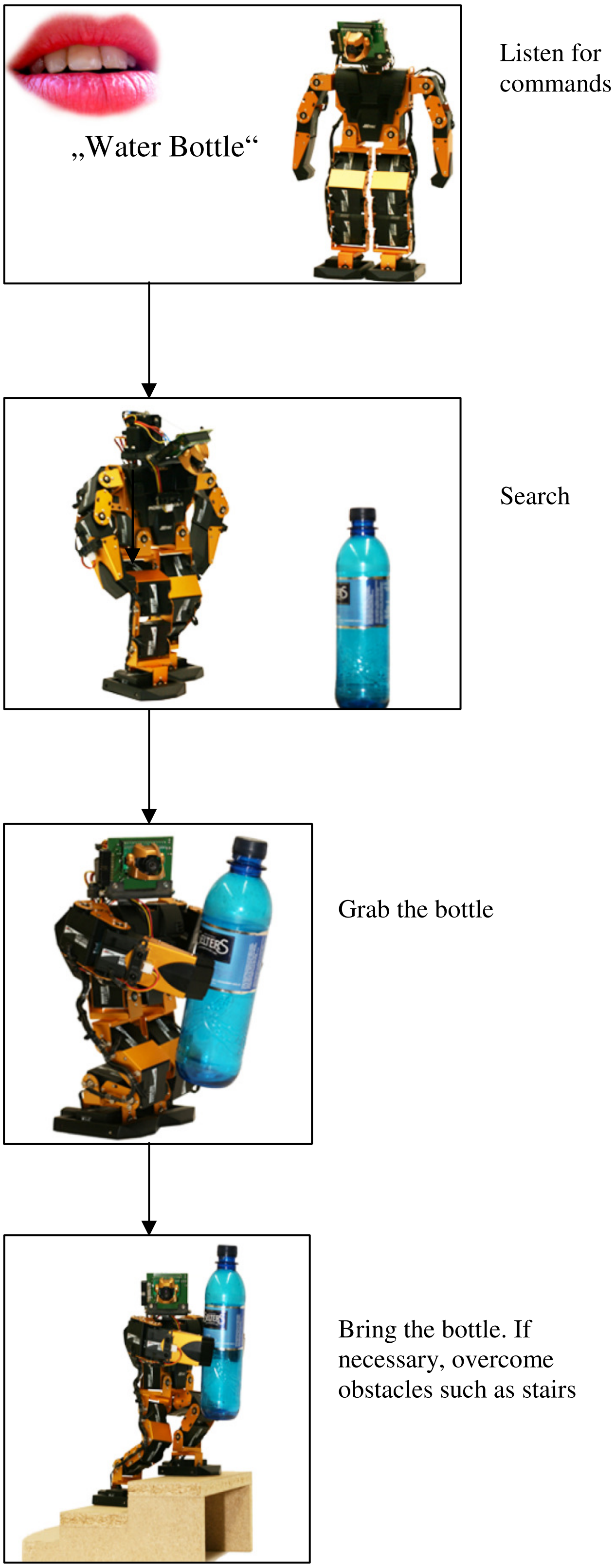
„Water Bottle“
Listen for commands
Search
Grab the bottle
Bring the bottle. If necessary, overcome obstacles such as stairs

Figure 3. Object Search and Fetch

VII. Swarm Robots

If the robot sensors (feed back signal of motor positions, tilt sensor) indicate that an object is too heavy to grab or to maintain a stable position it is set to the state "need help". This triggers the transmission of a message to other robots via Bluetooth which then walk to their "colleague" in order to assist by grabbing and carrying the object together (Fig. 4) as swarm robots [10].

If a robot has found an object it is not instructed to grab, it informs other robots about the object type and position. A robot which has received the goal to fetch this particular object can go directly to it without the need for searching by itself.

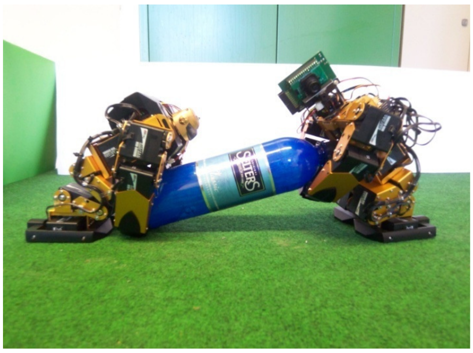

Figure 4. Swarm Robots grabbing a bottle together

VIII. Self Generating Will

The robot so far obeys the user and executes the task straightforward, but it cannot cope with conflicting requirements and has no behavioural intelligence [2]. The limitation of the robot presented above is the dependence of the state $s_j\left(\underline{x}\right)$ on the task-related sensory input $\underline{x}$, only. By weighting the speech and object recognition input by experiences made by a weighting factor w_j, a sort of feelings with respect to a given state could be implemented which will asses complex situations better. Adding the sensory input $\underline{d}$ for acquiring data about dangerous situations such as a high temperature and drives such as hunger (low battery status) as well as the desire for praise by the user for having achieved a goal, the robot can prevent actions which would harm it. Hence, the states

$$s_j\left(\underline{y}\right) = s_j\left(\left(\underline{x}^T W, \underline{d}^T\right)^T\right)$$

depend on the feelings $\underline{x}^T W$ and drives $\underline{d}$ where the diagonal matrix W contains the weighting factors w_j in the diagonal. In order to avoid that experiences once made dominate the robot's behaviour forever, a time component is introduced which reduce the weighting factors as a function of time. A drive "adventure" could cause the robot to walk around in order assess new situations if it is in the state "idle", i.e. when it has not received any instruction by the user. A quality of state function [8]

$$Q\left(s_j\left(\underline{y}\right), a_i\right)$$

is introduced to calculate a measure how "desirable" a state is. The generation of the will to perform an action

$$a_i\left(s_j\left(\underline{y}\right), g, a_{i-k}, \underline{y}\right)$$

is not rule based but the result of optimizing the quality of state function

$$Max\left\{Q\left(s_j\left(\underline{y}\right), a_i\right)\right\} \rightarrow a_w$$

This operator predicts the quality of all states $s_j\left(\underline{y}\right)$ which can be reached from the current state with respect to the possible actions a_i that can be executed from the current state and selects the maximum. A quality criterium could be the sum of all components of $\underline{y}$, i.e. the feelings and drives. As a result the robot generates the will to perform an action a_w which maximizes the quality of state.

VII. Summary

An humanoid robot has been developed which understands spoken instructions and can act accordingly by recognizing the environment with intelligent sensors.

Acknowledgments

The research was funded by the Fachhochschule Frankfurt a.M. and the Gesellschaft für technische Weiterbildung e.V.

References

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer Verlag, 2008

[2] G. Doeben-Henisch, Humanlike Computational Learning Theory. A Computational Semiotics Perspective, IEEE Africon, 2009

[3] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Prentice Hall, 2008

[4] K. Hirai, M. Hirose, Y. Haikawa, T. Takenaka, The Development of Honda Humanoid Robot, IEEE Int. Conf. Robot. Autom.,pp.1321-1326, 1998

[5] Tae-Seok Jin, Bong-Kee Lee and Jang Myung Lee, "AGV Navigation Using a Space and Time Sensor Fusion of an Active Camera", International Journal of Navigation and Port Research, Vol 27, No 3, 2003

[6] Peter Nauth, Embedded Intelligent Systems, Oldenbourg Verlag , 2005

[7] Bruno Siciliano and Oussama Khatib, Handbook of Robotics, Springer Verlag, 2008

[8] Kristen Stubbs and David Wettergreen, "Anatomy and Common Ground in Human-Robot Interaction: A Field Study", IEEE Intelligent Systems, March 2007

[9] Ron Sun, "Cognitive Social Simulation Incorporating Cognitive Architectures", IEEE Intelligent Systems, September 2007

[10] V. Trianni, S. Nolfi, M. Dorigo, "Cooperative hole-avoidance in a swarm-bot", Robot. Auton. Syst. 54