THE PROTOTYPE OF HUMAN – ROBOT INTERACTIVE VOICE CONTROL SYSTEM

Miroslav Holada, Igor Kopetschke, Pavel Pirkl, Martin Pelc, Lukáš Matela, Jiří Horčička and Jakub Štilec
Faculty of Mechatronics and Interdisciplinary Engineering Studies, Technical University of Liberec
Hálkova 6, Liberec, Czech Republic
{miroslav.holada, igor.kopetschke, pavel.pirkl, martin.pelc, lukas.matela, jiri.horcicka, jakub.stilec}@tul.cz

Keywords: Distributed speech recognition, robot control, image processing.

Abstract: This contribution shows our recent activities of human – robot voice-control system. The keynote lays in design of a system that allows easy control and programming of various robots by a uniform interactive interface. Generally the robots are actuated by sets of control commands, sometimes by a manual control interface (touchpad, joystick). The operator has to know the control commands, syntax rules and other properties necessary for successful robot control. Our system offers commands like "move left" or "elevate arm" that are translated and sent into the corresponding device (robot). Speech recognition is based on an isolated-word HMM engine with distributed recognition system (DSR). The vocabulary may contain thousands of word and short phrases. It allows us to design an easy dialogue scenario for the robot control system. The system allows direct robot control like moving or programming short sequences. A video camera is included for displaying the working place and employed algorithms of image processing allow the system to detect the position of objects in the working area and facilitate robot's navigation.

1 INTRODUCTION

The goal of this project is a PC-software-based interactive system for general-purpose robot voice-control. This paper describes the designed prototype, its structure and the dialogue strategy in particular.

The interactive control of robots could be used in special situations, when a robot is working in dangerous areas and no programming beforehand is possible. It could also be used in a situation when supervised learning for robot's later autonomous operation has to be done, without knowledge about the robot's programming language. The presented paper follows on this reality.

2 PROJECT FEATURES

The project is based on a former research. The research involved a voice-control dialog system, speech recognition vocabulary design and speech synthesis feedback for user command confirmation. Together with a scene manager and a digital image processing module, it forms the core of the control system as shown in figure 1. The components are described below.

2.1 Scene Manager

The scene manager forms a connection between the main program (engine) and the image processing part. It actually controls the image processing module and initiates image acquisition and processing. Using the processed image data, it updates the scene database, keeps track of objects exposed on the scene and provides the scene object and image data to the main engine. It is also aware of the robot's coordinate system and plans the robot's movement when requested by the engine.

The database itself consists of two types of data. It contains the list of parametrized objects detected on the scene as well as the robot calibration data. The latter allows mutual image-space to robot-space coordinate translation which is used in robot navigation. Each object detected on the scene is internally represented as a data object (class instance), all the objects are stored in a dynamic list. Some of the attributes are: a unique object identifier, object's shape descriptor, central point coordinates,

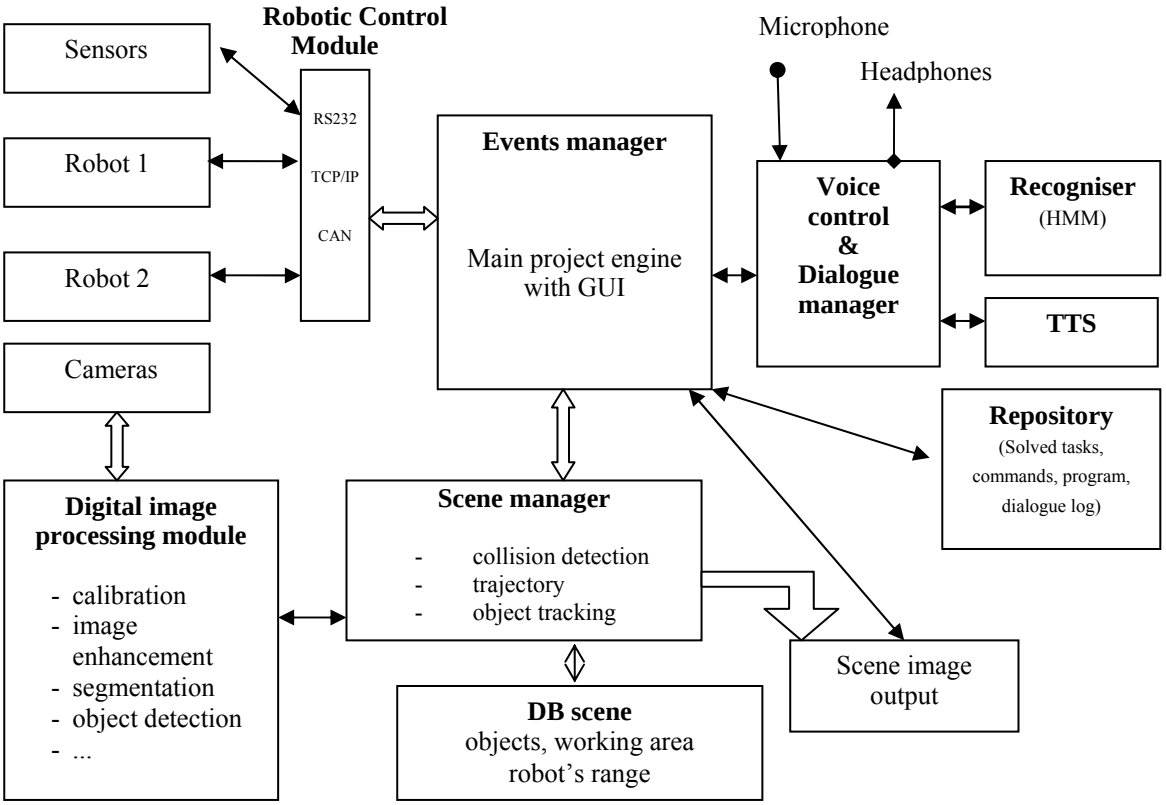

Figure 1: Scheme of the designed system.

bounding rectangle etc. Such data allows smooth object manipulation and serves as a base for object collision avoidance along the manipulation trajectory.

The scene manager also combines unprocessed camera image with scene data to highlight detected objects and to present them to the user via a GUI as shown in figure 2. The user has a view of the computer's scene understanding and may correctly designate objects of interest in his or her commands.

Being in its early stages, the project currently works only with 2D data and relies on the user's z-axis navigation aid. The system is expected to incorporate a second camera and 3D computer vision in the future to become fully 3D aware.

2.2 Image Recognition

The robot's working area is captured by a colour high-resolution digital camera (AVT Marlin F-146C, 1/2'' CCD sensor). The camera is placed directly above the scene in a fixed position. We implemented a simple interactive method to synchronize the robot's coordinate system (XY) and the camera's one using pixel units. We prepare modifications to compensate geometric distortions introduced by a camera lens.

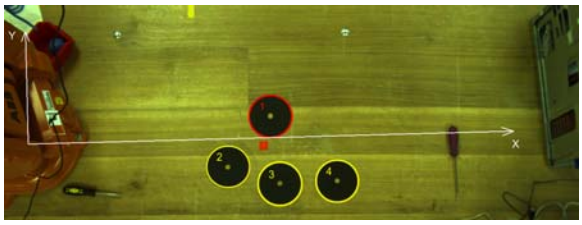

Figure 2: Working scene with highlighted objects of circular shape and robot's coordinate system.

Digital image processing methods are placed in a DIP library which is served by the scene manager with the object database. Figure 2 shows the circular object detection using the reliable Hough transform (HT). HT is commonly used for line or circle detection but could be extended to identify positions of arbitrary parametrizable shapes. Such edge-based object detection is not too sensitive to imperfect input data or noise. Using a touch-display or verbal commands it is possible to focus the robot onto a chosen object (differentiated by its color or numbering) and than tell the robot what to do.

2.3 Robots Description

The prototype system uses a compact industrial general-purpose robotic arm (ABB IRB 140). The robot is a 6-axes machine with fast acceleration, wide working area and high payload. It is driven by a high performance industrial motion control unit (S4Cplus) which employs the RAPID programming language. The control unit offers extensive communication capabilities - FieldBus, two Ehternet channels and two RS-232 channels. The serial channel was chosen for communication between the robot and the developed control system running on a PC.

Figure 3: Robots used.

The robotic control SW module simplifies the robot use from the main engine's point of view. It abstracts from the aspects of physical communication and robot's programming interface. It either accepts or refuses movement commands issued by the core engine (depending on command's feasibility). When a command is accepted, it is carried out asynchronously, only notifying the engine once the command is completed.

2.4 Distributed Computation

Most of the system's modules are developed and run on a standard PC to which the robot is connected. Since some of the software modules require significant computational power, the system's response time was far from satisfactory when the whole system ran on a single computer. Therefore, the most demanding computations (namely the object recognition and the voice recognition) were distributed to other (high performance) computers via network (TCP connections).

3 DIALOGUE STRATEGY

The voice interface between an operator and the controlled process is provided by a speech recogniser and a text-to-speech synthesis (TTS) system (both for Czech language). The TTS synthesis system named EPOS was developed by URE AV Prague. It allows various male or female voices with many options of setting.

The speech recognition is based on a proprietary isolated word engine that was developed in previous projects. The recogniser is speaker independent, noise robust, phoneme based with 3-state HMM (Hidden Markov Models) and 32 Gaussians. It is suitable for large vocabularies (up 10k words or short phrases) and allows us to apply various commands and their synonyms.

The dialog system is event-driven. We can categorize the events into three fundamental branches: operator events, scene manager events and device events.

3.1 Operator Events

Operator events usually occur in response to operator's requests. For example commands which are supposed to cause robot's movement, object detection, new command definition or detection of a new object. This kind of event can occur at any time, but the dialog manager has to resolve if it was a relevant and feasible request or if it was just a random speech recognition error.

Although the acoustic conditions in robotic applications usually involve high background noises (servos, air-pump), the speech recogniser works commonly with 98% recognition score. If the operator says a wrong command or a command out of context (for example, the operator says "drop" but the robot doesn't hold anything) then the manager asks him or her for a feasible command in the stead of the nonsensical one.

3.2 Scene Manager Events

This sort of event occurs when the scene manager detects a discrepancy in the scene. For example when the operator says "move up" and the robot's arm moves all the way up until the maximum range is reached. When this happens a scene event is generated and the system indicates that the top position was reached.

Other scene event occurs when the operator wants to take up an object, but the system does not know which one because of multiple objects

detected on the scene. This event generates a query to the operator for proper object specification.

3.3 Device Events

These events are produced by external sensors and other components and devices connected to the system. They are processed in the event manager where corresponding action is taken. The response manifests itself in the form of a request for the operator, or more often causes a change in robot's behaviour.

The difference between scene manager events and device events is that scene events are generated by the system itself (based on a known scenario, robot geometry, object shape and position). They are computed and predictable. On the other hand, device events' time cannot be exactly predicted before they actually happen.

3.4 Example of Dialog

For a simpler robot orientation and navigation the positions on the scene are virtualized. They are named by Greek letters like "Position α" or "Position β". These virtual positions may be redefined to suit the operator's needs. A blind-area may also be defined and it is completely omitted from any image processing and anything in this area is completely ignored.

We can gather up black disks (see figure 2.) and put them to some other place on the scene. This place is defined like "Position α" and we setup the "blind area" on the same coordinates. After that the operator starts a dialog, for example:

```
   "Start recording new command."
   "I'm recording"          ...robot says
   "Search objects"
   "I'm searching ... Four objects were
found"
   "Move on first"
   "I'm moving ... Done"
   "Take it."
   "Ok"
   "Move on position alpha."
   "I'm moving ... Done"
   "Put it"
   "Ok"
   "Stop recording."
   "I stop the recording. Please, say
new command"
   "Search" "Disks" "Done"
   "New command is entered and named:
Search disks. Is it right."
   "Yes"
   "Repeat command"
   "Enter command"
   "Search disks"
   "OK"
   ....
   "No object found. Repeating done."
```

The robot finds the remaining three disks and puts them in the selected area. If any disk cannot be found then the robot interrupts executing the given command and waits for next action.

4 CONCLUSIONS

The system is especially usable as an accessory robot control interface for assistant and second-rate operations. The presented prototype now cooperates only with one industry robot (ABB) but the robotic control module may easily be extended to support other robots (Katana, mobile robots, etc.) as well.

The designed system offers robot control and robot task programming even to people without explicit programming language knowledge. It is sufficient for the operator to know the Czech voice interface of the presented system.

ACKNOWLEDGEMENTS

This work was supported by the internal grant IG FM 2007/002.

REFERENCES

Nouza, J., Nouza, T.: A Voice Dictation System for a Million-Word Czech Vocabulary. In: *Proc. of ICCCT* 2004, August 2004, Austin, USA, pp. 149-152,

Holada, M.: The experiences and usability of distributed speech recognition system DUNDIS. In: *Proc. of 14th Czech-German Workshop „Speech Processing"*, September 2004, Prague, Czech Republic, pp. 159-162,

Hanika, J, Horak, P.: Text to Speech Control Protocol. In: *Proc of the Int. Conf. Eurospeech'99*, Budapest, Hungary, September 5-9, 1999, Vol. 5, pp. 2143-2146.

Šonka, M., Hlaváč, V., Boyle, R. D., *Image Processing, Analysis and Machine Vision.* PWS, Boston, USA, second edition, 1998.