Adaptive Multimodal Interaction with Industrial Robot

Milan Gnjatović, Jovica Tasevski, Milutin Nikolić, Dragiša Mišković, Branislav Borovac, Vlado Delić
University of Novi Sad, Faculty of Technical Sciences, Novi Sad
milangnjatovic@yahoo.com, {tasevski, milutinn, dragisa, borovac, vdelic}@uns.ac.rs

***Abstract*— This paper reports a spoken natural language dialogue system that manages the interaction between the user and the industrial robot ABB IRB 140. To the extent that the dialogue system is multimodal, it uses three communication modalities: (i) spoken language (automatic speech recognition and text-to-speech synthesis), (ii) visual recognition of the figures and determination of their positions, and (iii) typed text. To the extent that the dialogue system is adaptive, it takes the verbal and spatial contexts into account in order to adapt its dialogue behavior and to process spontaneously formulated user commands of different syntactic forms without explicit syntactic expectations. The industrial robot is slightly modified and enabled to manipulate over graphical figures, following the instructions of the dialogue system.**

I. INTRODUCTION

In most human-robot interaction scenarios, the robot performs a preset operation after a voice command is uttered. In addition, these scenarios rarely include an industrial robot (cf. [1], [2], [3]).

This paper reports a spoken natural language dialogue system that manages the interaction between the user and the industrial robot ABB IRB 140. The interaction scenario lies within the scope of therapeutic interaction between the language therapist and the child with receptive language difficulties (cf. [4], [5]). During the interaction, the user and the system share two related contexts – a spatial context and a verbal context. The spatial context consists of a set of wooden three-dimensional figures that differ in:

- base – triangle, circle, square, and rectangle,
- height – thin (7mm) and thick (15mm),
- size – big and small,
- color – red, blue, and yellow.

These figures are often used in therapy for receptive language impairment. At the start of the interaction, they are placed randomly on a table. Both the user and the system can manipulate these figures.

In the verbal interaction, the user utters instructions that relate to manipulation of the figures (e.g., pointing to a figure, or shifting a figure). The user is allowed to spontaneously utter commands of different syntactic forms (including elliptical, minor and context-dependent commands), and the task of the system is to properly interpret these verbal stimuli and to perform the requested operations. In cases when the user does not provide all necessary information, the system applies an adaptive dialogue strategy intended to help the user to specify the required information.

The system integrates the four subsystems:

- The visual subsystem implements the functionalities of visual recognition of the three-dimensional figures placed on the pad, and determination of their positions.
- The audio subsystem implements the functionalities of automatic speech recognition and text-to-speech synthesis.
- The cognitive subsystem implements the functionalities of natural language processing and adaptive dialogue management.
- The external robotic subsystem is the industrial robot ABB IRB140, an anthropomorphic robotic arm with six degrees of freedom.

The following sections introduce these subsystems in more detail. The related work is discussed for each subsystem separately.

II. THE VISUAL SUBSYSTEM

The task of the visual subsystem is to recognize the figures on the pad, and to determine their positions. It uses two "AXIS 211" IP-based network cameras in a stereoscopic configuration. They are positioned approximately 70cm from the figures placed on the pad. The stereoscopic configuration enables the subsystem to acquire the depth information for each figure, i.e., the information of whether the given figure is thin or thick.

Imperfections of the optical system cause a distortion of the acquired image. Therefore, in order to remove the image distortion, the both cameras are calibrated. In addition, during the calibration, the relative position of the cameras is acquired. This information is necessary for the purpose of rectification and determination of the disparity map. After calculating the disparity map, the information on the relative position between the cameras is used to calculate positions of each image pixel in three-dimensional Cartesian space, the origin of which is located at the center of the left camera.

All the figures are placed on a flat paper pad of size A4. The origin of the corresponding coordinate system is positioned at the center of paper, so that the x-axis is parallel to the long edge of the pad, and the y-axis with the short edge of the pad. The significance of the pad is twofold. With respect to digital image processing, it represents a referent coordinate system for determining the positions of the figures. With respect to the spatial context, it represents a working frame of the robot. In other words, two independent modules – the robot and the image processing module – share the same coordinate

frame, which makes the mapping between the figure positions and the robot acts significantly less difficult.

The vision subsystem also comprises the software modules implemented in the programming language C++, based on the open source library of programming functions for real-time computer vision "OpenCV 2.3.1" [6]. The pad position is detected during the initialization phase. For this purpose, four rectangular markers are positioned in the corners of the pad, and it is assumed that the pad remains stationary. Then the rectification is performed, and the transformation matrix between the disparity map and the coordinate system attached to the pad is calculated.

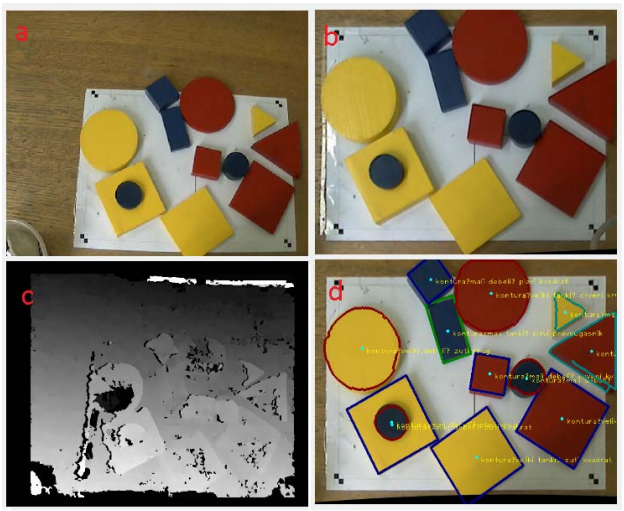

Figure 1. Digital image processing: (a) image created by the left camera; (b) image after perspective normalization; (c) the disparity map; (d) detected object representation.

After completing the initialization phase, the raw images are captured by the cameras, as given in Fig. 1(a). In the next step, the perspective is normalized, so that the image plane coincides with the pad's plane, as shown in Fig. 1(b). In this perspective, the figure positions in the image are linearly related to the figure positions in the coordinate system of the pad. In addition, the three-dimensional figures are perceived as projected onto a two-dimensional plane (e.g., a cylinder is projected to a circle, a square cuboid is projected to a square, etc.) which simplifies their classification to a great extent.

In the segmentation phase, the Canny edge detector was applied to find the edges and to determine the closed contours. The descriptors used to classify the object shapes include: contour area, contour perimeter, the smallest enclosing circle, and minimum area bounding rectangle. Based on these descriptors, the shapes of the figures (circle, triangle, square or rectangle) are determined. The size of the figures (small or large) is determined based on the contour area. The color of the figures (red, blue or yellow) is determined based on the mean color value. The thickness (thick or thin) of the figures is determined based on the disparity map, given in Fig. 1(c), and the transformation matrices created during the initialization phase.

It should be noted that the processes of perspective normalization and segmentation are independent of the processes of calculating the disparity maps and spatial coordinates of the figures. Therefore, in order to improve the performance of the visual subsystem, these processes are parallelized. For this purpose, the library "Intel Thread Building Blocks" [7] was applied. The overall result of the visual recognition is stored in a vector of recognized figures. Each figure in this vector is represented by a structure containing the following information: size, shape, color, thickness, and x and y coordinates of the figure centre in the pad coordinate system.

III. The Audio Subsystem

The audio subsystem implements the functionalities of automatic speech recognition (ASR) and text-to-speech synthesis (TTS). This subsystem expands upon previous work in the field of speech technologies for the Serbian language [8].

The TTS engine is developed as a standalone SAPI 5 speech synthesizer, called *anReader*. It includes several male and female synthesized voices, and allows the pronunciation adjustment to a considerable extent.

The ASR engine is based on the *Alfanum* automatic speech recognition system. It is a continuous speech recognizer involving small- and medium-sized vocabularies. The system is speaker-independent and phoneme-based, with a three-state hidden Markov model (HMM). An elementary HMM is a triphone model, representing a phoneme in a particular left and right context. It is important to note that the application designer is allowed to introduce an arbitrary set of words (i.e., a vocabulary) at the initialization time. A vocabulary is defined by a grammar that describes all utterances that may be produced by the user in the given interaction domain. To illustrate this, a simplified grammar, in Backus-Naur form, is provided, describing a potential verbal command in the interaction between the user and the robot introduced in this paper:

```
Command    = [[$action] [$color]
               [$figure] [$direction]];
Figure     = TRIANGLE | SQUARE |
             CIRCLE | RECTANGLE;
Color      = RED | BLUE | YELLOW;
Action     = SHOW | MOVE;
direction  = UPWARDS | DOWNWARDS |
             LEFTWARDS | RIGHTWARDS;
```

Thus, an example of a valid command would be: *shift*$_{[action]}$ *red*$_{[color]}$ *triangle*$_{[figure]}$ *upwards*$_{[direction]}$. The important limitation of this approach is that the user is forced to follow the preset grammar. In the reported study, this grammar-related restriction is relaxed. In order to increase the level of naturalness of the spoken interaction, the users are given only a set of keywords and key-phrases that relate to entities in the spatial context. The system does not introduce any syntactic expectation. This is illustrated by the following definition of a command:

```
command  = [{$words}];
words    = TRIANGLE | SQUARE |
           CIRCLE | RECTANGLE |
           RED | BLUE | YELLOW |
           SHOW | MOVE |
           UPWARDS | DOWNWARDS |
           LEFTWARDS | RIGHTWARDS;
```

The user is allowed to utter his commands more flexibly, e.g., "shift red triangle upwards", "red triangle, shift upwards", etc. However, he is also allowed to utter elliptical and context-dependent commands, or even semantically incorrect commands (e.g., "show rightwards leftwards"). It is the task of the cognitive subsystem to appropriately interpret such inputs. The next section discusses this research question in more detail.

The speech recognition engine is integrated into the ASR IP-based server. This enables a remote access to the server, so that the application designer needs to develop only a relatively simple software client. In addition, a specialized library that provides a set of functions for communication between PC applications and the ASR engine (i.e., the *callback* mechanism) was developed.

One of the functions in this library is to collect samples from the PC microphone. After silence detection (i.e., the moment when the user has already finished his utterance), the embedded algorithm sends a recognition request to the server and waits for a response. A schematic representation of the communication protocol between the client and the server applications is given in Fig. 2. The server's response consists of two arrays. The first array comprises the recognized words. The second array comprises numeric values, each of which represents the recognition reliability for a corresponding word. This information (i.e., the recognized command) is then forwarded to the cognitive subsystem for an interpretation.

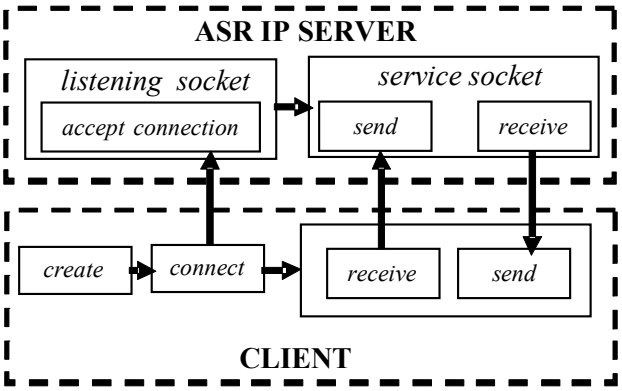

Figure 2. A schematic representation of the communication protocol.

IV. THE COGNITIVE SUBSYSTEM

This section briefly discusses and illustrates (cf. Fig. 3) the cognitive subsystem that implements the functionalities of natural language processing and adaptive dialogue management.

References [4], [9] introduce and exemplify a model of attentional state in human-machine interaction. It draws upon and integrates neurocognitive understanding of the focus of attention, and notions of attention in computational and corpora linguistics. To the extent that the model is computationally appropriate, it addresses the research problem of robust automatic processing of different syntactic forms of spontaneously uttered user commands with no explicit syntactic expectations. Therefore, this model underlies the implementation of the cognitive subsystem.

User$_1$:	*Give me that angular figure* *(In Serbian: Дај ми тај угласти предмет)*
System$_2$:	The formulation of you command is too general. I can interpret it in 18 different ways. Please be more specific. *(In Serbian: Формулација Ваше команде је превише уопштена. Могу да је разумем чак на 18 начина. Молим Вас да будете конкретнији.)*
User$_3$:	Move the red thin square downwards *(In Serbian: Помери црвени танки квадрат на доле)*
System$_4$:	The system forwards the command to the robotic system.
User$_5$:	The circle thick blue downwards *(In Serbian: Круг дебели плави на доле)*
System$_6$:	The system forwards the command to the robotic system.
User$_7$:	The yellow square *(In Serbian: Жути квадрат)*
System$_8$:	How should I interpret your command: thin yellow square, or thick yellow square? *(In Serbian: Како да интерпретирам Вашу команду: танки жути квадрат или дебели жути квадрат?)*
User$_9$:	Thin *(In Serbian: Танки)*
System$_{10}$:	You may choose among the following actions: show thin yellow square, and move thin yellow square. *(In Serbian: На располагању су Вам следеће акције: покажи танки жути квадрат, и помери танки жути квадрат)*
User$_{11}$:	I want you to show it *(In Serbian: Желим да га покажеш)*
System$_{12}$:	The system forwards the command to the robotic system.
User$_{13}$:	And now upwards *(In Serbian: А сад на горе)*
System$_{14}$:	The system forwards the command to the robotic system.

Figure 3. Dialogue fragment.

However, the existing model is extended with a dialogue strategy. It may be described as follows: If the user does not completely formulate a command (and the system cannot recover the omitted information from the context) or utters an ambivalent command (i.e., it can be assigned more than one interpretation in the given context), the system takes the initiative in the interaction and guides the user to complete the started command, or to resolve the ambiguity, by stating iterative questions.

Fig. 3 shows a dialogue fragment that took place during the testing of the cognitive subsystem, using typed text as a communication modality. It illustrates the processing of the user commands and the dialogue strategy. The dialogue fragment is translated into English. Descriptions

of non-verbal actions performed by the system are underlined.

V. The Robotic Subsystem

The industrial robot ABB IRB 140 is an anthropomorphic robotic arm with six degrees of freedom. It is equipped with IRC5 robot controller, which, in this case, has only RS232 serial communication. The software ABB Rapid [10] is used for programming the robot.

The robot was slightly adopted by positioning a flexible plastic stick with a rubber layer on its top at the end of the robot's arm. The rubber layer is used to increase the friction between the stick and the figures on the pad. Thus, the robotic system does not need a gripper to shift a figure over the pad, or to point to the figure on the pad. A flexible stick was used in order to prevent accidental damages of the figures, the pad, etc.

VI. Integration

The visual, audio and cognitive subsystems are integrated within a PC application. The visual subsystem forwards information on figures that are detected on the pad to the cognitive subsystem. Similarly, the audio subsystem forwards the recognized user's command to the cognitive subsystem. Taking these two information sources and the context of the interaction into account, the cognitive subsystem tries to process the user's command. If the command can be processed (i.e., the cognitive subsystem has all necessary information), its code is forwarded to the robotic system that performs the requested operation. Otherwise, if the command is incomplete, ambivalent or semantically incorrect, so it cannot be processed, the cognitive subsystem generates an appropriate linguistic content (e.g., asking the user to reformulate his command, etc.) and forwards it to the text-to-speech module within the audio subsystem.

A schematic representation of the integrated system is given in Fig. 4. The integrated system is given in Fig. 5.

VII. Conclusion

This paper reports a spoken natural language dialogue system that manages the interaction between the user and the industrial robot ABB IRB 140. To the extent that the dialogue system is multimodal, it uses three communication modalities: (i) spoken language (automatic speech recognition and text-to-speech synthesis), (ii) visual recognition of the figures and determination of their positions, and (iii) typed text. To the extent that the dialogue system is adaptive, it takes the verbal and spatial contexts into account in order to adapt its dialogue behavior and to process spontaneously formulated user commands of different syntactic forms without explicit syntactic expectations. The industrial robot is slightly modified and enabled to manipulate over graphical figures, following the instructions of the dialogue system.

The future work includes integration of the dialogue system with an anthropomorphic humanoid robotic arm, more complex manipulation over the figures, and introduction of more advanced adaptive and assistive dialogue strategies.

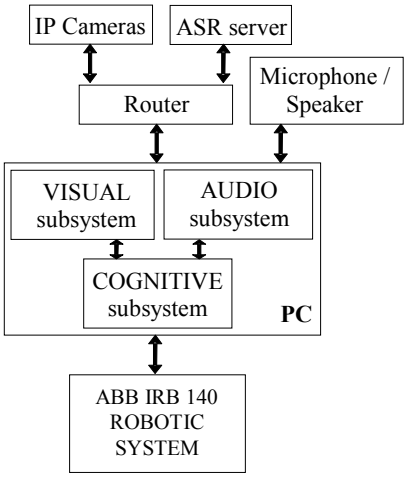

Figure 4. A schematic representation of the integrated system.

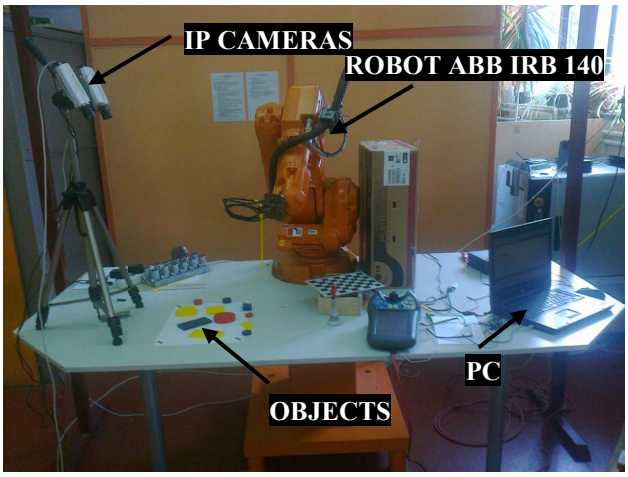

Figure 5. Integrated solution.

Acknowledgment

The presented study is performed as part of the projects "Design of Robots as Assistive Technology for the Treatment of Children with Developmental Disorders" (III44008) and "Development of Dialogue Systems for Serbian and Other South Slavic Languages"' (TR32035), funded by the Ministry of Education and Science of the Republic of Serbia. The responsibility for the content of this paper lies with the authors.

References

[1] J. N. Pires, "Robot-by-voice: Experiments on commanding an industrial robot using the human voice", Industrial Robot: An International Journal, Vol. 32, 2005

[2] J. N. Pires, G. Veiga, R. Araujo, "Programing-by-demonstration in the coworker scenario for SMEs", Industrial Robot: An International Journal, Vol. 36, 2009

[3] D. Rambabu, R. Nagaraju, B. Venkatesh, "Speech Recognition of Industrial Robot", International Journal of Computational Mathematical Ideas, Vol. 3, 2011

[4] M. Gnjatović and V. Delić, “Attention and linguistic encoding of motion events in human-machine interaction”, in Selected papers from the 3rd International Conference of Syntax, Phonology and Language Analysis, S. Halupka-Rešetar, M. Marković, T. Milićev, and N. Milićević, Eds.vCambridge Scholar Publishing, in press.

[5] Srđan Savić, Miloš Jurošević: “Design Of Modular Robot Arm With 7 Degrees Of Freedom”, Procedings of ETRAN, Zlatibor, Serbia, 11-14 June, 2012, (In Serbian: Razvoj robotske ruke antropomorfnih karakteristika)

[6] R. Laganière, “OpenCV 2 Computer Vision Application Programming Cookbook”, Birmingham, Pakct Publishing 2011.

[7] J. Reinders, “Intel Thread Building Blocks”, Birmingham, O’ Reilly 2007.

[8] V.Delić, D. Pekar, R. Obradović, N. Jakovljević, D. Mišković : “A Review of AlfaNum Continuous Automatic Speech Recognition System”, XII international conference “Speech and Computer” (SPECOM’2007), Moscow, Russia, 15-18. october, 2007

[9] M. Gnjatović, M. Janev, and V. Delić, “Focus tree: Modeling attentional information in task-oriented human-machine interaction”, Applied Intelligence, 2011, 10.1007/s10489-011-0329-5. [Online]. Available: http://dx.doi.org/10.1007/s10489-011-0329-5

[10] “ABB Rapid reference manual”, ABB Automation Technologies AB, Robotics, 2005