Human-natural communication with "smart-room" by combination of various modalities

Bc. Jakub Rajniak, prof. Ing. Gregor Rozinaj, PhD.
Slovak University of Technology, Ilkovičova 3, 821 19, Bratislava, Slovakia
09rajniak@gmail.com, gregor.rozinaj@gmail.com

Abstract. **This document deals with the issue of device management in smart laboratory smart-room by human natural communication with computer based on gestures and speech.**

Keywords **- Smart-room, speech recognition, gesture recognition, Kinect**

I. INTRODUCTION

Nowadays, the most typical ways to operate with the computer are touchscreens or keyboard and mouse. Computer responds to us by output devices, especially screens or speakers. In principle, all I / O devices are designed to make a connection between the computer device and the external environment so as to enable it to interact with the computing device. It is important to note the fact that both end elements are located in two different systems. A man in an analogue world and a computer device in a digital world where everything goes on binary system. The research and development of I / O devices continues in the spirit of adapting the devices to users. It is easy for a person to understand visual or audio output from a computer, but it is more complicated for the computer, generally trying to prevent a misunderstanding of inconspicuous communication. The aim is for the computer to understand the communication that one learns from his natural nature through his two basic senses: hearing and sight.

Senses

A. Computer hearing

Hearing is an essential sense for verbal spoken communication. In developing area of speech recognition, it is necessary to take into account the different parameters of the voice that the computer does not take as distinct from the human being as obviosity. Researchers work primarily on recognizing a particular person, content, what has been said, and even what has been said to be determined by the speech prologue. The non-verbal audio language of the language in which the interrogation fits in can also be included here. For the computer, the audio signal in this area is not very different because they are recorded by microphones that give the same parameters only with different values [1].

B. Computer sight

Up to eighty percent of all perceived information is obtained by the sight of man. In this area, there is a wide range of possibilities for the development of data processing obtained from the space in front of the scanning device. The main task of the device is to capture data to distinguish the human or human part from the surrounding environment and then evaluate whether or not the person wants to communicate. Types of alternative visual non-verbal communication with a computer are, for example, the following:

- Gesture processing - Gestation is the most perceived form of non-verbal communication, usually only complements and confirms the spoken word, but there are situations where it carries full meaning equal to the spoken word. Devices dealing with this issue are, for example, Kinect [2] or Leap.
- Mimic processing - Mimic is an area in which a person gives his feelings. These feelings can be evaluated and understood by the device as commands [3].
- Eye View Processing - Eye View Direction also has its communication value in the communication sphere, and it can be combined with deliberate glowing [4].

II. THE "SMART-ROOM"

Intelligent rooms in the world as well as the smart-room at our Institute of Multimedia Information and Communication Technologies offer natural communication to humans with various electrical or electronic devices. The traditional form of room control via switches, mechanical buttons or remote controls remains, but this concept is made up of an innovative superstructure. Smart-room can handle voice commands or gesture commands. Smart-room recognize who is in the room by the face or voice, but only if person is in the user's database of the room. Personal settings such as music, lighting, or switching to different profiles, such as a visit profile, can be assigned to the user. The second specialty of smart rooms is a network connection between devices. Communication takes place on the network layer via ports and IP addresses. In the room is a central computer that processes and evaluates data from sensors such as Kinect v2. If the application positively evaluates the command, it sends the message either directly to specific devices or to a device capable of generating an infrared signal for devices with an infrared receiver. Network devices implemented in the smart-room laboratory at the Institute for Multimedia Information and Communication Technologies are Quido ETH 8/8 and GC-100-12. Both of these devices were used to communicate with end electrical devices.

Figure 1. "Smart-room" Laboratory

III. PROCESING OF BIOMETRIC INPUTS

Process of natural communication between man and computer is not done with push buttons, mouse or keyboard, but with gesture and speech commands. In the hierarchy of the proposed architecture, the Kinect v2 is a sensor for audio and visual inputs from the user. Subsequently, these data are processed in a programmed software application. The created program is based on application examples from the software development toolkit Kinect v2 offered by Microsoft for Kinect developers. The program can track the human skeleton and recognize speech, which are the conditions of natural communication. By combining these two ways, we can create a condition for sending messages to network modules to turn on or off an electrical device. The main principle of natural object control in a smart-room laboratory is to point to the object with the right hand and a voice command to define the event or the state of the electrical device that we want to achieve.

Figure 2. Gesture - pointing at lamp

A. Pointing gesture

The basic gesture of pointing at something is the most natural and simple gesture in commanding people in human communication, so it is good for the computer to detect it. This gesture creates an imaginary line in the space that can be created with two points. We created a function for parametric expression of the line where the input parameters are the x, y, z of the elbow joint and the right hand tip from sensor Kinect v2. There is no continuous line in the digital world, and it is unnecessary to specify a high density of points that would generate a pseudo-continuous line for the program's calculation speed. For optimizing and simplifying the evaluation, we straighten the line to the relative dimensions of the room.

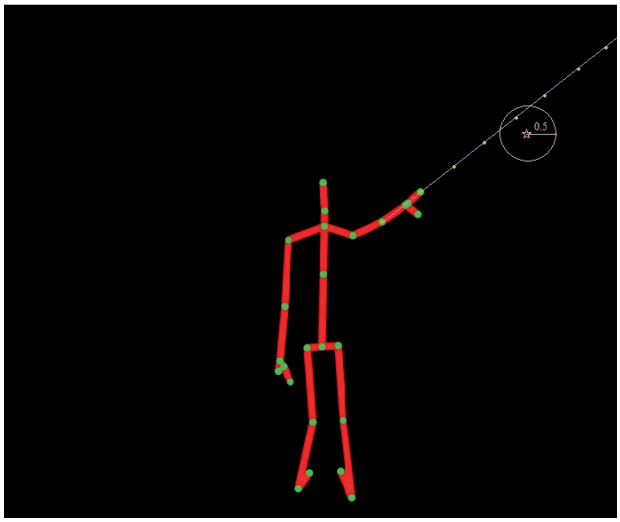

Figure 3. Pointing line

B. Speech commands procesing

Speech processing runs through the Microsoft Speech Recognition namespace, helping us to create an application to access and modify, word, phrase, or word-finding real-time algorithms. The first condition was to create a Grammar object, consisting of sets of rules and constraints that define words or phrases. The Grammar application uses the input to evaluate the meaningful verbal commands. Using the grammar class constructor, we created a grammatical object from the file, but it can also be programmatically created using the GrammarBuilder and Choices classes. An XML file contained a simple text structure to define words where the dictionary could be represented by a word combination, one word, or a combination of multiple words. Instances of SpeechRecognitionEngine provide access to installed recognizers (Runtime Languages) to perform speech recognition. A Runtime Language includes the language model, acoustic model, and other data necessary to provision a speech engine to perform speech recognition in a particular language.

Part of code is fusing detection of simple pointing gesture and speech command. It creates complex condition for detecting of conscious behavior and helps to avoid accidental detection that could be made in system that operate just with gestures or just with speech recognition.

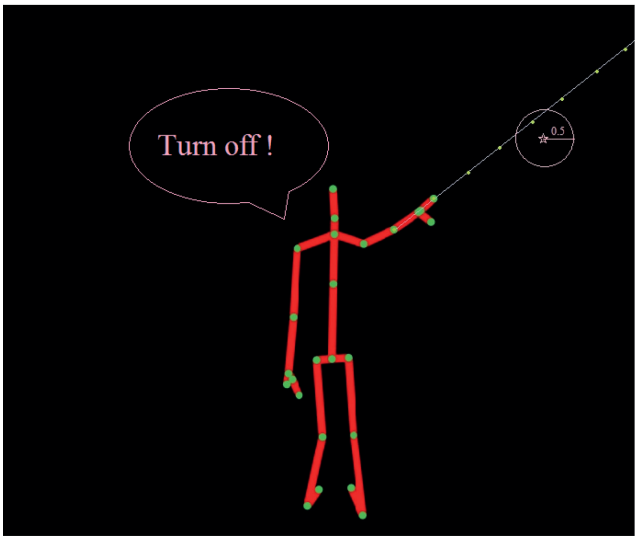

Figure 4. Combined condition for detection

IV. ARCHITECTURE SETUP

Software application consist of two main parts that holds all project together. Fist part of code is fusing detection of simple pointing gesture and speech command. Also in this part is defined location of devices in laboratory smart-room relative to position of the sensor Kinect v2. Second part of code send messages to network modules via Ethernet. Network modules located in laboratory smart-room are Quido ETH 8/8 and Global Cash-100-12. Quido uses relay components that open or close the circuit of ceiling lighting or electric sockets. Electric sockets are good for devices that operate between two states on and off. Global Cash-100-12 is module that can communicate with other devices by infrared signals.

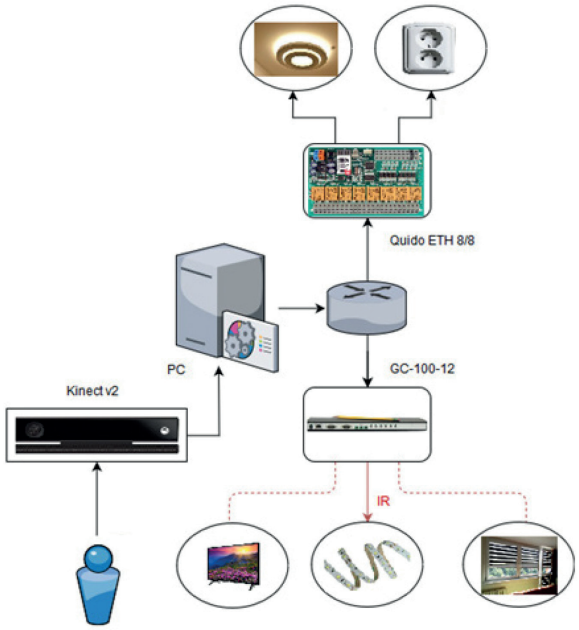

Figure 5. Architecture of devices in smart-room

ACKNOWLEDGMENT

The research described in the paper was financially supported by the H2020 project NEWTON, No. 688503 and VEGA project INOMET, No. 1/0800/16.

REFERENCES

[1] "Voice recognition software" 17 April 2018 [Online] Available: http://www.explainthatstuff.com/voicerecognition.html

[2] "Kinect Learn, Kinect for Windows" 22 April 2018 [Online] Available: http://go.microsoft.com/fwlink/?LinkId=247735

[3] Depth Sensor Shootout 13 April 2018 [Online] Available: https://stimulant.com/depth-sensor-shootout-2/

[4] "What Is Eye Tracking?" 22 April 2018 [Online] Available: http://www.tobii.com/tech/technology/what-is-eye-tracking/