electronics

MDPI

Article

A Multimodal User Interface for an Assistive Robotic Shopping Cart

Dmitry Ryumin [1], Ildar Kagirov [1,*], Alexandr Axyonov [1], Nikita Pavlyuk [1], Anton Saveliev [1], Irina Kipyatkova [1], Milos Zelezny [2], Iosif Mporas [3] and Alexey Karpov [1,*]

1 St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 199178 St. Petersburg, Russia; ryumin.d@iias.spb.su (D.R.); axyonov.a@iias.spb.su (A.A.); antei.hasgard@gmail.com (N.P.); antoni-fox@ya.ru (A.S.); kipyatkova@iias.spb.su (I.K.)

2 Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, 301 00 Pilsen, Czech Republic; zelezny@kky.zcu.cz

3 School of Engineering and Computer Science, University of Hertfordshire, Hatfield, Herts AL10 9AB, UK; i.mporas@herts.ac.uk

* Correspondence: kagirov@iias.spb.su (I.K.); karpov@iias.spb.su (A.K.)

Received: 2 November 2020; Accepted: 5 December 2020; Published: 8 December 2020

Abstract: This paper presents the research and development of the prototype of the assistive mobile information robot (AMIR). The main features of the presented prototype are voice and gesture-based interfaces with Russian speech and sign language recognition and synthesis techniques and a high degree of robot autonomy. AMIR prototype's aim is to be used as a robotic cart for shopping in grocery stores and/or supermarkets. Among the main topics covered in this paper are the presentation of the interface (three modalities), the single-handed gesture recognition system (based on a collected database of Russian sign language elements), as well as the technical description of the robotic platform (architecture, navigation algorithm). The use of multimodal interfaces, namely the speech and gesture modalities, make human-robot interaction natural and intuitive, as well as sign language recognition allows hearing-impaired people to use this robotic cart. AMIR prototype has promising perspectives for real usage in supermarkets, both due to its assistive capabilities and its multimodal user interface.

Keywords: assistive robotics; service robotics; multimodal user interface; sign language processing; gesture interface; speech recognition; voice interface

1. Introduction

Assistive robots are robots that help to maintain or enhance the capabilities usually of older persons or people suffering from functional limitations. There is a vast discussion concerning the necessities of older people, and assistive robots can surely cover some of them [1–9]. Thus, assistive robots help people with injuries to move and to maintain a good social life status, resulting in psychological and physical well-being. The prime example of this strategy is provided in the EQUAL project [10], aimed at enhancing the quality of life of people suffering from moving disability. The developers conducting the EQUAL project propose a number of steps leading to shopping facilitation. Among the steps are the development of assistive mechanized shopping cart, software, and improved infrastructure supporting people with a moving disability. Assistive robots and, more broadly, assistive technologies are designed to support or even replace the services provided by caregivers and physicians, reduce the need for regular healthcare services, and make persons who suffer from various dysfunctions more independent in their everyday life. Assistive robots often have a multimodal interface that facilitates

human-machine interaction. Most often, such robots are mobile and have access to wireless computer networks, which allows them to be used as telepresence robots, facilitating continuous communication with other people. In some cases, a robot is equipped with a robotic arm and can move relatively light objects to help the user.

According to works [11,12], assistive robots fall into two major categories: rehabilitation robots and socially active robots. The former is designed to provide mainly physical assistance, while the latter function as personal assistants or service robots, mainly improving the psychological well-being of the user.

This article presents a human-machine interface for controlling the prototype of assistive robotic platform AMIR (assistive mobile information robot), some aspects of which have been previously described in papers [13,14]. The AMIR project has been developed by the St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS, http://hci.nw.ru/en/projects/17) since 2018 as an assistive robotic shopping cart for supermarkets and food stores. Among the main features of AMIR are the contactless user-cart interaction possibility via a gesture and voice-based user interface, Russian sign language recognition, and a high level of autonomy (route tracking, navigation inside a supermarket, and providing information about food products and their location in the store). AMIR has been developed to assist people who suffer from progressive hearing loss, as well as different groups of people who need assistance in supermarkets (e.g., elders). The aim of this work is to describe in detail the architecture of the interface of the robotic platform used for the interaction with the user, as well as the scope and application perspectives of the current prototype in the context of assistive robotics.

The remainder of this article is structured as follows. An overview of assistive and service robotic shopping carts is provided in Section 2. The architecture of the robotic platform is described in Section 3. In Section 4, we present the architecture of the human-machine interaction interface integrated into the AMIR robotic platform. In Section 5, preliminary experiments of speech and gesture recognition are presented. Finally, in Section 6, conclusions and perspectives on the presented implementation as well as future work directions are given.

2. Related Work

Robotic shopping assistants are mainly used in navigation tasks to address user's shopping needs/interests, usually defined by a shopping list. Navigating through a supermarket in search of products is not always easy for a customer. Most often, finding the needed department or aisle takes most of the time that customers spend in a store. The use of robotic platforms can solve this problem, saving the customer's time and energy.

Different aspects can be considered in the classification of robotic shopping assistants. The choice of classification criteria depends on the research task pursued by researchers. In our work, we focus on the human-machine interface and interaction, and the literature review is driven by the functions supported by robotic platforms.

2.1. Carts that Follow the Customers

As most people want a faster shopping process, several approaches have been proposed in which robotic assistants and robotic carts follow the customer. In work [15], the authors presented a robotic transport system to help customers. This system consists of a guide robot, a cart robot, and cameras. The guide robot is an autonomous mobile robot with a localization function, followed by the cart robot. The cameras are used to detect obstacles and people around the robot. The obstacle detection system uses 32 sets of ultrasonic sensors connected in series. In addition, the platform has a second detection system consisting of three sets of laser rangefinders.

A similar robotic shopping cart system was presented in work [16], where the cart provides information about the goods and also helps in shopping. In order to start shopping, the customer needs to log on to the robot's system. After that, the customer has to provide information about the desired

products. The robot will suggest shopping mall recommendations and build a route. In the event the customer logs out of the system, the cart can return to the starting point in an autonomous mode. The device records customers' purchases for further analysis of customers' preferences. The computing system is located at the base of the cart, as well as software-compatible devices. The robot has a laser scanner on the front for the detection of obstacles. In order to receive information about products, the robot is equipped with a radio frequency identification (RFID) tag reader. These tags are also used to localize the cart. Interaction with the robot is performed via a touch screen tablet.

A rather interesting development of an assistive cart is presented in the paper [17]. This article presents a shopping cart with the ability to autonomously follow a customer while he/she is shopping. The development is aimed at helping people who find it difficult to push the cart due to any physical dysfunction (injury, pregnancy, disability, or aging). Most customers spend a lot of time looking for the products they need; in this regard, the authors propose a system for finding the shortest route based on the customer's shopping list. The shortest route to the required aisles/shelves is calculated by a genetic algorithm (GA) [18] using the traveling salesman problem (TSP) [19] model. The customer uses the mobile app on his/her smartphone for the creation of a shopping list. This list is then sent to the server, where the route that the robot will follow is generated. The cart moves to the first item in the list. After the cart has reached the point of destination, it waits for the customer to pick up the item and mark it in the mobile app. The cart performs this action for each item in the shopping list. The robot also uses Microsoft Kinect for mapping, localization, and navigation.

Paper [20] proposed a design of a shopping assistant robot using deep learning technologies. Information about the environment is sent to the robot from two Kinect v2 sensors. The first sensor helps the robot move and localizes the environment, while the second sensor recognizes and tracks the customer. The authors presented several use cases: (i) Kinect 1 detects surroundings using a depth map to identify passages and shelves; (ii) Kinect 2 detects people using skeleton detection and tracks faces; (iii) Tracking customer during driving. While the robot is moving, thanks to sensor 1, sensor 2 must track the customer. In addition, the speed of the robot adapts to the customer's pace; (iv) Tracking customer emotions with Kinect 2. Emotion is a kind of feedback about the service.

2.2. Mobile Phone-Based Customer-Robot Interaction

In [21,22], the authors presented an assistant robotic system designed to help older and disabled people by providing information about products and calculating the total amount in the cart. The user connects to the system via a mobile phone. After connecting, the user must be identified by a personal identification number (PIN), which is registered in the system. The robot finds the best route in the supermarket, subject to the customer's preferences. The customer can correct the automatically proposed route if some goods are not on his/her list. If the user decides his/her own route, then the robot follows him/her.

2.3. Robotic Cart for Elderly and People with Special Needs

Elders and disabled people also need assistance with shopping. Unfortunately, most of them cannot go shopping on their own. Usually, in this group of people, they have caregivers who accompany and help them in shopping. In this regard, caregivers have to spend a lot of time and effort. In work [23], the authors focused on assisting elderly customers in shopping. A robotic shopping cart with shopping information is presented. Experiments were performed in a real store. The robot performs autonomous movements in designated directions, as well as determines its location. Besides, the robot has a switch for manual control. It is noted that older people have difficulty in using the touchpad, and in this regard, the authors proposed to reproduce information about the product by sound. Purchase history is planned to be used to suggest products next time, as well as the ability to update the location map.

The smart shopping cart can also be used outside the store, for example, to deliver groceries and accompany the elderly. The authors of the paper [24] presented the CompaRob assistant robot. The device is based on an Erratic-Videre mobile robotic platform equipped with a grocery basket.

This platform has an integrated PC, laser rangefinder, and other sensors. The robot weighs 13.5 kg, and the payload is up to 20 kg. CompaRob contains three lead-acid batteries for 2 h autonomy and ultrasonic sensors. The robot assistant follows the customer by reading signals from an ultrasonic ring attached to the person's leg.

In [25], a mobile robot assistant called Tateyama is presented. This robot is designed for moving and lifting a shopping cart. The mobile platform for controlling the robot is equipped with two cameras, three sets of wheels (front, middle, and rear) for climbing stairs, and two manipulators with six degrees of freedom for holding the cart. Remote control of the robot is performed using a game controller. The cart has a hand brake mechanism. This mechanism contains two shafts that are used to press or release the cart brakes. In addition, the authors developed a step-by-step stair climbing method for the robotic system. The robot performs human-like movements.

2.4. People with Visual Impairments

Robotic systems can be used to assist people with visual impairments. An informative overview of mobile technologies for people with visual impairment is provided in [26]. The authors classified assistive solutions into three categories: tag-based systems (usually, RFID and NFC tags), computer vision-based systems, and hybrid systems. Examples of tag-based systems and approaches are given in works [27,28]. In paper [29], the development and application of robot assistant RoboCart are described. This platform looks like a hand cart. The robot's hardware contains radio frequency identification (RFID) tags, a platform microcontroller, and a laser rangefinder. RFID tags can be attached to any product or clothing and do not require an external power source. In addition to this, it is a simple and cheap solution. The device software consists of three components: a user interface, a route planner, and a behavior manager. The route planner and the behavior manager partially implement spatial semantic hierarchy [30]. Following it, information about space is divided into four levels: control level, causal level, topological level, and metric level. RoboCart has the two following disadvantages, namely difficulty in turning around in aisles and limited spatial sensing (50 cm from floor level), making difficult the detection of billboards installed on shelves.

Computer vision-based systems identify objects without RFID or NFC tags, directly utilizing information about features of the objects. The disadvantage of this approach is that additional devices are often prerequisite for the system to function. The paper [31] introduced a design for smart glasses used to assist people with visual impairments.

Hybrid systems combine strong points from both approaches. For example, in works [32], a smartphone camera is used to identify QR-codes on product shelves and RFID tags to navigate through a store.

2.5. The Next Group Includes Robotic Platforms, the Key Feature of Which Is Remote Control and Remote Shopping

People tend to exceed their budget when they are shopping in large stores. They also end up in long queues after shopping to pay for their purchases. The smart shopping cart helps solve such problems. These devices can automatically count the contents of the carts. In this regard, the authors of papers [33,34] proposed a robotic system developed for remote shopping. In this platform, a control element is a manipulator with two degrees of freedom. The manipulator consists of four suction cups designed to grip and hold objects of different textures, shapes, and masses. The customer has Internet access to the robot and has the ability to control shopping. This is possible using video captured from a camera mounted on the robot. According to test results, the positioning error of the robot relative to an object does not exceed 15 mm. The robot adds one product to a basket every 50 s. Specifically, 30 s are required for selection and scanning, and 20 s to transfer the selected product to the basket. However, the proposed robotic system has important disadvantages—for example, a limited application area (only fruits and vegetables).

In paper [35], the use of the light interactive cart 3S for smart shopping was proposed. The prototype of the cart was created by encapsulating off-the-shelf modularized sensors in one small box, fixed on the handle. This solution can be regarded as a lightweight or is regarded so by the authors. 3S consists of wireless routers, a shopping cart, and a management server. As a rule, products of the same type are usually placed on the same aisle/shelf, so a wireless router was installed on each shelf to be able to correlate the type of product and its location. Once the router detects the arrival of a customer, the system understands what products the customer is looking for. In addition, to help with the selection of goods, the smart cart is able to search for a convenient route and calls a store employee if the customer moves along the same path many times or the cart does not move for a long time. According to the results of the experiment, the 3S cart saves about 10–25% of the time spent on customer's navigation in the sales area if compared to the pathfinding algorithm A* [36].

Today, automated counting of cart contents is gaining popularity. In this regard, the authors of [37] offered an intelligent shopping cart. The cart has a system that calculates and displays the total cost of products put into it. The customer can pay directly for his/her using the cart. This solution lets the user skip the process of scanning products at the checkout and significantly saves his/her time. In paper [38], the authors presented an intelligent shopping system using a wireless sensor network (WSN) to automatize invoice processing. Article [39] demonstrated the successful use of an ultra-high frequency (UHF) RFID system mounted on a smart shopping cart for the same purpose.

In papers [40,41], an anthropomorphic robot assistant named Robovie was presented. The robot performs three main tasks: localizing people and identifying and tracking faces. Robovie has two actuators with four degrees of freedom, a robotic head with three degrees of freedom, a body, and a mobile wheel-type base. There are two cameras and a speaker attached to the head, and a wide-angle camera and a microphone on the shoulder. The developed robotic system can recognize the customer's gender; identify people using RFID tags; give information about purchases during communication, and provide navigation along the route using gestures. The robot has partial remote control. This is necessary for avoiding difficulties with speech recognition.

2.6. Issues Can Arise When Multiple Robots Are Functioning in the Shopping Room at Once

In this case, control and robot-to-robot interaction system are required. The system of four robots Robovie, described in [42], consists of three components: a task manager, a route coordinator, and a scenario coordinator. The task manager distributes tasks between robots based on their location and human behavior. The path coordinator generates routes for the movement of robots based on information about their location. The scenario coordinator provides communication between devices. Six laser rangefinders for localizing people and robots are installed. In [43], the authors proposed using a group of robots designed to distribute advertising coupons in the shopping room. The system consists of two similar robots, the main difference of which is only in their external characteristics. Both robots can do spoken dialogue interaction with customers and can print coupons. The first anthropomorphic robot with a height of 30 cm is based on Robovie-miniR2, having two limbs with eight degrees of freedom and a head with three degrees of freedom. The second 130 cm height humanoid robot is equipped with two actuators and a robotic head. The head has a speaker, a camera, and a microphone. For the implementation of interactive behavior, corresponding modules are used. These modules control the robot's speech, gestures, and non-verbal behavior in response to human actions. The behavior selector controls the robot's behavior using pre-designed sequence rules and sensor inputs. The authors developed 141 behavior scenarios and 233 episodes with four types of behavior classes: route control (101 scenarios), providing store information (32 scenarios), greeting (seven scenarios), and coupon printing (one scenario).

Based on the above-mentioned platforms, it can be seen that the idea of assistive robotic carts implementation is not new and has been of high importance for a considerable period of time. Despite the active use of different contactless ways of interaction of the user with the robotic cart, none of the platforms described above makes use of a gesture interface. Combination of gesture and

speech modalities are found even more rarely, with no previous study found in the literature for Russian sign language recognition by robots. Currently, the only Russian sign language recognition/synthesis system that has gained popularity is Surdofon [44], which combines software solutions and online service tools for the translation of Russian sign language into Russian-sounding language. However, Surdofon does not actually recognize gestures (except for the online service, with human signers engaged in translation): a deaf or a user with hearing disabilities has to use the textual modality in order to input information, while the tool answers in spoken Russian, which is converted into sign form using the application. Efforts are being made to develop assistive robots, which can interact with the user via sign languages [45–47], but none of them combine assistive technologies. At the same time, supporting people with hearing impairments using assistive technologies is of significant importance. Only in Russia, according to the 2010 census, more than 120 thousand people were using sign language to communicate in their everyday life. Moreover, the use of gesture modality substantially expands the possibilities for human-machine interaction when referring to socially specific, everyday signs and gestures.

3. AMIR Robot Architecture

AMIR robot has a modular architecture and an extensive set of sensors. It can be used particularly for human-machine interaction in a crowded environment, such as supermarkets and shopping malls. In this section, an overview of the design of the AMIR prototype is provided.

AMIR robot consists of two main units: the mobile robotic platform (MRP) and the informational kiosk (IK). MRP contains (1) a computing unit with Nvidia Jetson TX2/Xavier module, (2) wheelbase electric drives, (3) power supply block (44 Ah), (4) navigational equipment, and (5) interfaces for connectivity of the IK and peripheral equipment. MRP is essentially the driving unit of AMIR, and using the navigation equipment (lidars, obstacle detection sensors, nine-axis environment sensor MPU-9250), it performs navigation tasks: load transport, map composition, tracking a route, and following it, localization of AMIR in unknown environments. Lidars of RPLidar S1 model from SLAMTEC are used for the establishment of an interactive indoor map and device localization as well as for additional recognition of 3D objects around the robot, and laser sensors for obstacle detection are used for noise recognition and detection of obstacles, such as holes in the floor or small objects that are positioned lower than the reach of the lidar on the path of the platform. All the units of MRP are installed in an aluminum framework. Some photos of AMIR's general view are shown in Figure 1.

(**a**) (**b**) (**c**)

Figure 1. The assistive mobile information robot (AMIR) robotic platform with the informational kiosk (IK) mounted on the mobile robotic platform (MRP) as shown in: (**a**,**b**) side/front view with the interface, (**c**) the actual size, compared to an average height male person.

In more detail, the AMIR robot has the following design characteristics:

- Dimensions—60 × 60 × 135 cm
- Carrying capacity—20 kg
- Power supply unit—LiPo 44000 mAh 14.8V
- Omni-wheels (10 cm in diameter)
- 2 lidars with 360° sweep
- 16 obstacle sensors
- Computing unit with Nvidia Jetson TX2/Xavier.

An informational kiosk (IK) is a unit equipped with hardware as well as software and devices for human-machine interaction within a container block. IK unit contains modules responsible for human-machine interaction, such as a computing unit Intel NUC, wide-angle cameras, a touch screen, and the Kinect module. The IK information being displayed on the touch screen and obtained from the Kinect module is processed in the embedded computing unit Intel NUC. In turn, the computing units of MRP and IK communicate through a dedicated LAN connection. The interaction workflow of MRP and IK components is illustrated in Figure 2 below.

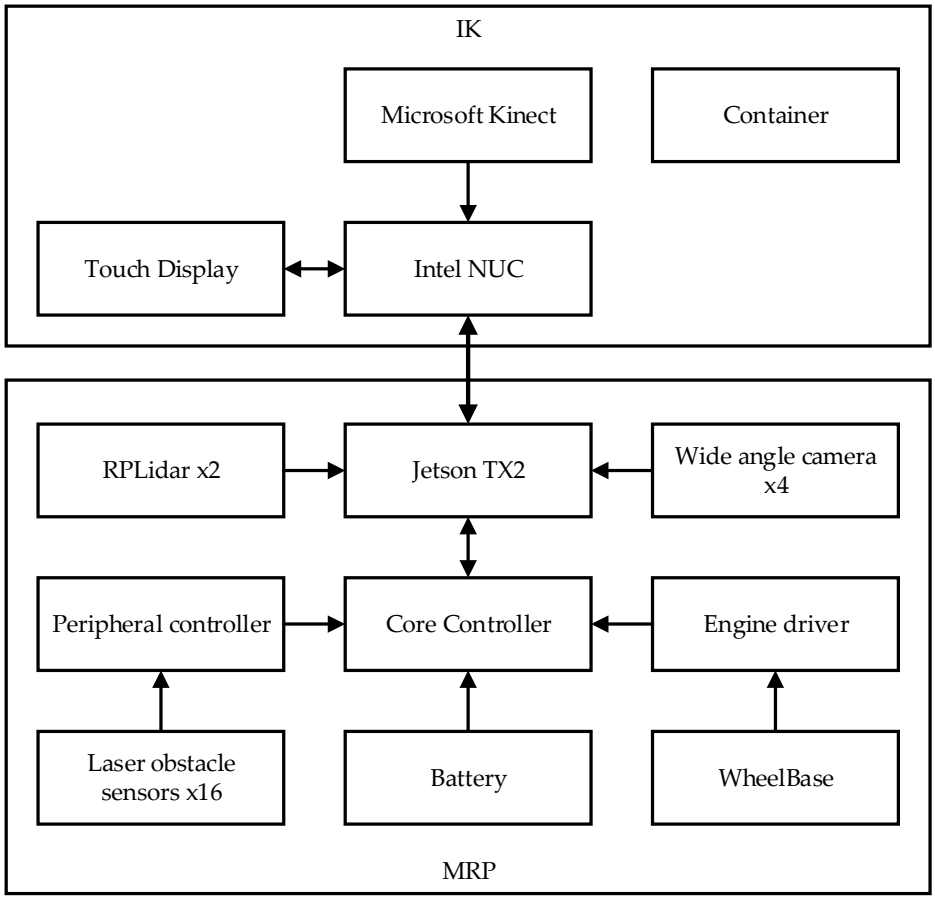


Figure 2. Block diagram of the architecture of AMIR robotic platform, consisting of two main units: the mobile robotic platform (MRP) and the informational kiosk (IK).

MRP core controller is an intermediate unit of data gathering and processing in the system. This controller performs low-level computation on an STM32F405VGT6 microprocessor and ensures connection with the peripheral devices using an SN65HVD233D CAN-transceiver. The main controller is connected to Nvidia Jetson TX2 via USB, and the controllers of peripheral devices and engine drivers are connected to Nvidia Jetson TX2 through CAN. In this controller, a nine-axis position sensor

MPU9250 is utilized. The feature of external indication is implemented using a 5V addressed LED strip, which utilizes the 1Wire protocol. Besides, optionally, additional devices can be connected to the AMIR architecture via idle ports GPIO and I2C.

To ensure the robot's navigation in unknown environments, the problem of simultaneous localization and mapping (SLAM) has to be solved. The SLAM problem consists of simultaneous detection of the condition of the sensor-equipped robot and mapping it with an unknown environment based on data obtained from these sensors. Path planning and localization modules (global planners) ensure the room mapping, localization, and path tracing to an intermediate target. Collision avoidance module (local planner) ensures platform motion to the intermediate target through an obstacle-free path, according to the global planner data, and by avoiding dynamic obstacles.

Lidar data in the form of sensor_msgs/LaserScan messages are published in ROS [48], as well as the odometry data, captured with Hall sensors and published as a nav_msgs/Odometry message. Furthermore, this information is utilized in the localization module for mapping, further transfer of mapping data into the motion planner for indoor navigation of the platform. The control system receives target instructions from the motion planner, and then it sends final instructions concerning motion velocity in the geometry_msgs/Twist message to the coordinate system of the platform.

Rao-Blackwellized particle filter approaches to SLAM, such as FastSLAM-2 [49,50], explicitly describe the posterior distribution through a finite number of samples—particles. Each particle represents a robot trajectory hypothesis and carries an individual map of the environment. Rao-Blackwellized particle filters reduce the number of particles required for estimation of the joint posterior of the map and trajectory of the robot through the factorization of this posterior. This factorization allows the computation of an accurate proposal distribution based on odometry and sensor data, which drastically reduces the number of required particles. In contrast to FastSLAM-2, where the map is represented by a set of landmarks, Grisetti [51] extends FastSLAM-2 to the grid map case. Efficient approximations and compact map representation presented in [52] significantly reduce computational and memory requirements for large-scale indoor mapping by performing necessary computations on a set of representative particles instead of all particles.

For room mapping, the localization module ensures the Gmapping package from the ROS framework. The Gmapping package implements the FastSLAM algorithm, which utilizes the particle filter to solve the SLAM problem. This filter allows the estimation of those parameters of the object that cannot be measured directly, deducing them from already known parameters. To assess the unknown parameters, the filter generates a set of particles, and each of them carries its own copy of the environment map. At the outset, all the particles are completely random, but at each iteration in the loop, the filter removes the particles that failed to pass the validation check until nothing, except the particles to remain, which are the closest to the true values of the parameters [51].

The software architecture for spatial navigation of AMIR is implemented with the ROS framework and is presented in Figure 3.

FastSLAM utilizes particle filters to assess the position of the robot and to map the environment. For each of the particles involved, the corresponding mapping errors are conditionally independent; therefore, the mapping process can be divided into a series of standalone tasks. The main objective of robot motion planning is to achieve a maximum velocity of motion to destination targets along the traced paths but in a completely collision-free manner. When solving this problem, secondary problems are occurred like the calculation of the optimum path, accounting for possible quirks in the execution of control instructions, as well as ensuring the fast generation of control instructions in the instances when unexpected objects appear in the dynamical environment the robot moves at.

To define a collision-free trajectory, the local planner of the navigation system utilizes the global dynamic window algorithm [53], aimed to achieve the maximum velocity of collision-free motion. The algorithm traces the path, using geometrical operations, provided that the robot traverses circular arcs and receives a control instruction (ν, ω), where ν is the velocity of straight motion, and ω is the velocity of rotational motion.

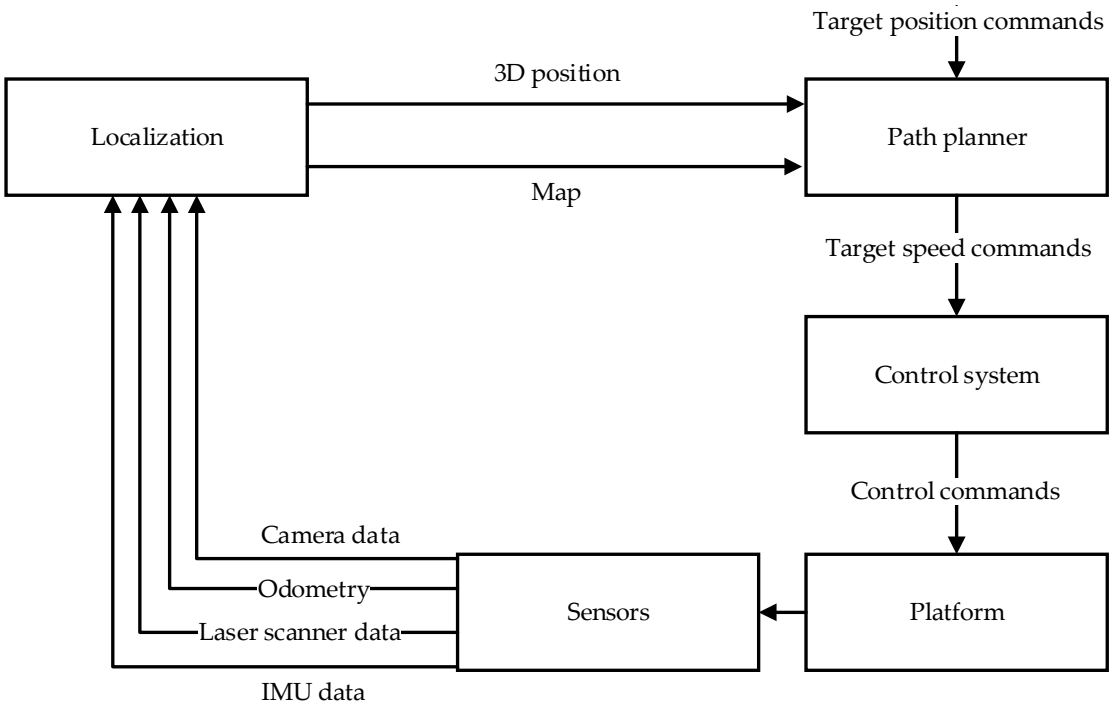


Figure 3. Software architecture for the navigation of AMIR robotic platform.

Among the advantages of the global dynamic window algorithm are the following ones: (1) fast reaction time, (2) moderate computing power required, and (3) collision-free motion path detection. The global dynamic window approach ensures the construction of high-velocity trajectories in unknown and dynamic environments.

The FastSLAM algorithm and the global dynamic window algorithm are successfully employed together in real-world models. The proposed software architecture, intended for robotic platform navigation in the real environment, ensures autonomous indoor mapping as well as path planning and obstacle avoidance. This is achieved using the information obtained from the sensors.

4. AMIR's Human-Machine Interaction Interface

The block diagram of the architecture of the user's interaction with the AMIR prototype is presented in Figure 4. The whole interaction process is carried out with a multimodal (touch, gesture, and speech) human-machine interface (MultimodalHMIinterface) software package.

The input data of the MultimodalHMInterface are video and audio signals. The Kinect v2 sensor is the device that receives the video signal (it is capable of receiving color video data and depth map). It calculates a 3-d map of the scene using a combination of RGB and infrared camera. The viewing angles are 43.5° vertically and 57° horizontally. The resolution of the video stream is 1920 × 1080 pixels with a frequency of 30 Hz (15 Hz in low light conditions). The inclination angle adjuster is pointed at changing vertical viewing angle within the range of ±27°. The color quality of the RGB video stream is 8 bits with a video stream resolution of 1920 × 1080 (Full HD) pixels and a frequency of 30 frames per second. The depth map can broadcast a transmitting video stream with a resolution of 512 × 424 pixels with 16 bits/pixel and at the same frame rate as an RGB video stream. For streaming the audio signal, a smartphone using the Android operating system is installed on AMIR. All above-mentioned receiving devices are installed on AMIR at a height between 1 and 1.5 m. The user performing interaction with AMIR has to keep the distance from the robot between 1.2 and 3.5 m. A smartphone-based application duplicates the touch screen mounted on the robotic platform prototype, allowing to switch modalities and navigate through menus.

Switching of modalities is performed through touch control, and the implementation of adaptive strategies is under development, i.e., in case of malfunction, the system will suggest the user switch to other interface channels. The implementation of automatic switching through voice or gesture interfaces will be possible at a production-ready level (Technology Readiness Level 7–8) of the robotic platform.

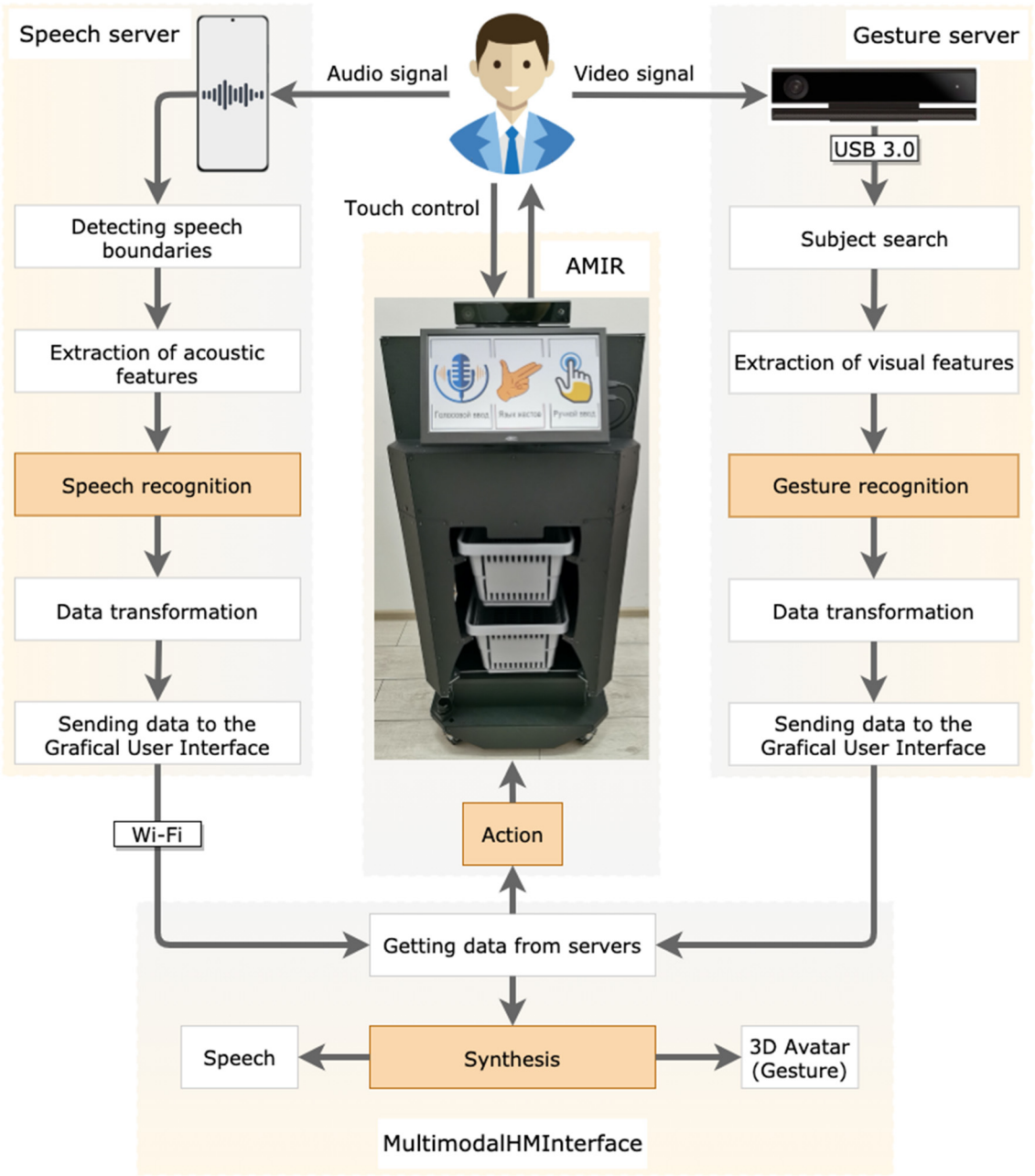
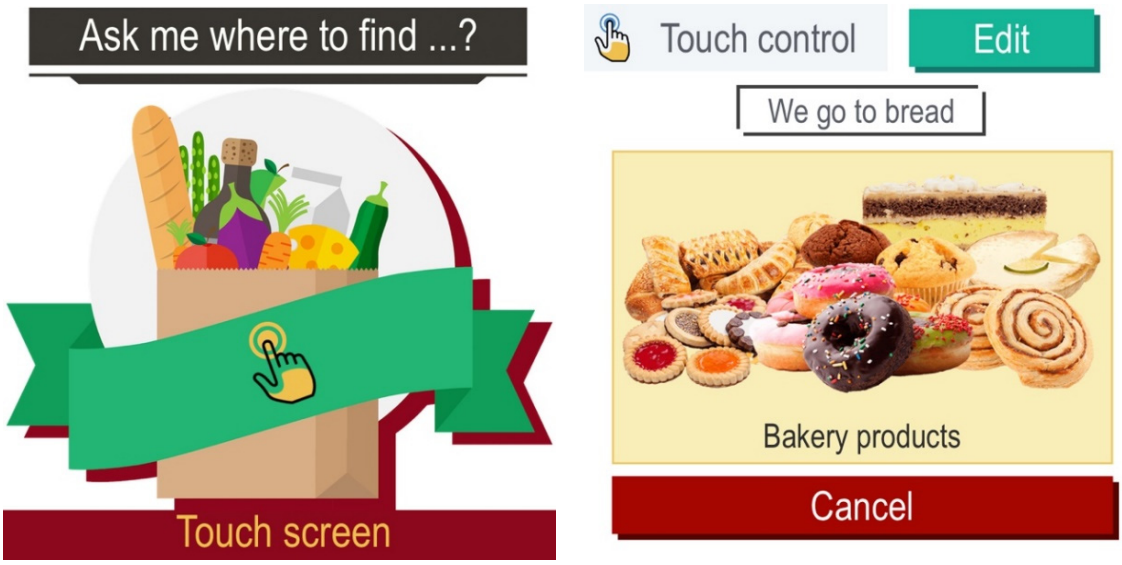


Figure 4. Block diagram of AMIR prototype user interface, with the use of 'MultimodalHMInterface' software package.

4.1. Touch Graphical Interface

The touch screen installed on the prototype AMIR allows the user to use the MultimodalHMInterface graphical user interface (GUI) through the touch modality, that is, a set of tools designed for user interaction with AMIR. The GUI of MultimodalHMInterface is based on the representation of objects and interaction functions in the form of graphical display components (windows, buttons, etc.). Therefore, the MultimodalHMInterface is currently an integral part of the AMIR prototype. Examples of screenshots from the GUI MultimodalHMInterface are presented in Figure 5.



(**a**) (**b**)

Figure 5. Examples of the graphical user interface (GUI) MultimodalHMInterface screenshots: (**a**) Start window; (**b**) Window for selecting a product from a specific category.

4.2. Voice Interface

Using voice for interaction is more natural for users than the usage of a graphical interface. Moreover, this type of interaction saves users time because pronouncing product names takes much less time than searching for it in the product list.

The implemented voice recognition technology is based on Android software; this solution increases the ease of use of the developed device. Besides, this solution simplifies the use of Android-based devices (one of the most common platforms) to interact with the system. The software of the voice interface of AMIR consists of a server and client parts. The server part is installed on an Android-based smartphone. The client part of the voice software runs on x64 computers with Microsoft Windows 8, 8.1, and 10 operating systems. The server part of the speech recognition system is installed on an Android smartphone because it uses the Google speech recognition API, which provides a set of tools for continuous speech recognition exclusively through the Android OS. For the purpose of automatic speech recognition, the open-source software from the Android OS is used to convert the audio/voice signal into a textual representation on Android running mobile devices [54].

The software carries out recognition of speech commands, the transformation of the recognized command to a digital constant (code), displaying it as a text on AMIR's monitor as well as pronouncing it by robotic/artificial voice via speech synthesis, by sending the code of recognized speech command to the control system of the AMIR.

In order to give a voice command to AMIR, the user should say a keyword or a phrase, which will be converted to a query for AMIR in order for AMIR to find the desired product or department of the shop. The query phrase may be arbitrary, but it should contain command words from the AMIR's dictionary in any grammatical form. Examples of such voice commands (both particular items and supermarket departments) used for interaction with AMIR robot are presented in Table 1.

The voice interface is based on voice activation technology, which means that an activation command is recognized in the speech stream. In order to reduce power consumption, the input audio stream is checked on the presence of speech. If speech is detected, the mode of activation command search is turned on. If the activation command matches a keyword, the search of command is performed on the speech signal stream after the keyword. The software sends the code of the recognized command to the IP address specified in the settings. The code of recognized speech command is sent to the control system of AMIR. The recognized command is displayed as a text on

the monitor and further generated as voice via speech synthesis technology. The average accuracy of recognized speech commands is above 96%.

Table 1. List of voice commands supported by assistive mobile information robot (AMIR).

Command (ID)	Category (Department)
yogurt (1), kefir (2), milk (3), butter (4), sour cream (5), cheese (6), cottage cheese (7), eggs (8)	milk products, cheeses, eggs
cake (101), cookies (102), bakery products (103)	confectionery
chocolate (201), candy (202)	chocolate products
long loaf (301), rusks (302), dried bread (303), bread (304)	bakery products
water (401), sparkling water (402), kvass (403), juice (404)	drinks
tomatoes (501), cabbage (502), cucumbers (503), potatoes (504), onion (505), carrot (506), oranges (507), apples (508), pear (509), lemon (510), bananas (511)	vegetables and fruits
Tea (601), coffee (602), pasta (603)	grocery
buckwheat grain (701), rice (702), oatmeal (703)	cereals
canned food (801)	canned food
salt (901), sugar (902), spice (903)	spice
sausages (1001), meat (1002)	meat
fish (1101), caviar (1102)	fish and seafood
sunflower oil (1201), yeast (1202), flour (1203)	vegetable oils sauces and seasonings
dumplings (1301), pizza (1302)	frozen semi-finished products
ticket window (1401), restroom (1402), output (1403)	departments and locations

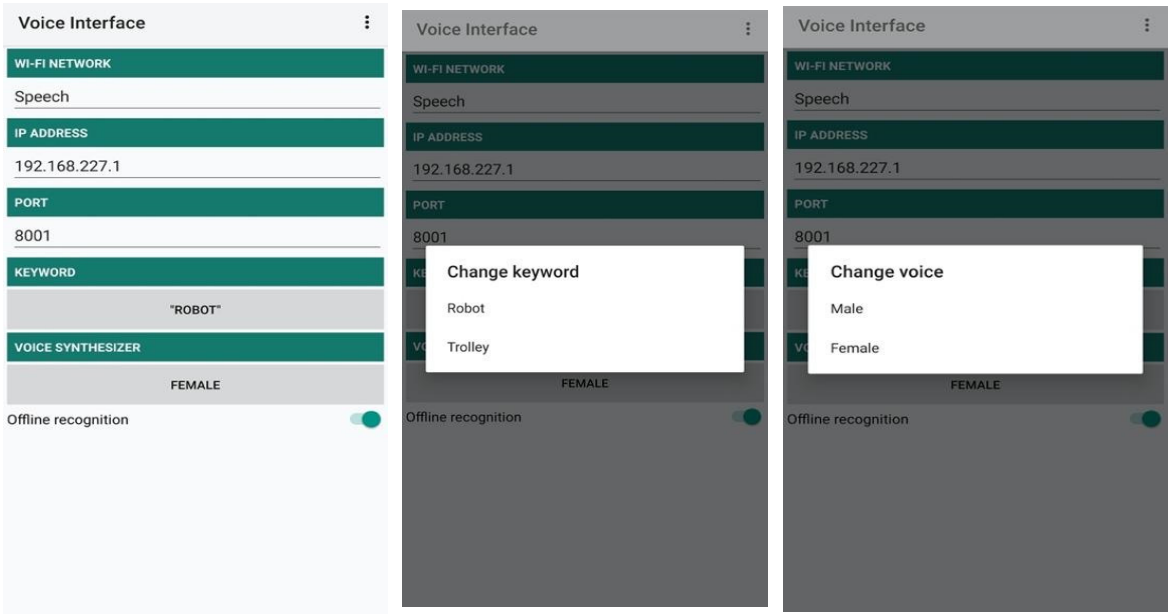
There are some start settings of the voice interface software. Before the first use of it, a Wi-Fi network connection should be established, as presented in Figure 6a. During the setting of a Wi-Fi connection, the user can set the network name, IP-address, and port. If the connection to the network is successful, a message "Connected to the required network" will appear. In the case of a second or multiple running of the software, a connection to the selected Wi-Fi network will be performed automatically.

The users can change a keyword. The default keyword is "Robot". In order to choose the keyword, the user should touch the corresponding field in the "Keyword" section. The menu for choosing the keyword is presented in Figure 6b. The user can set one of the following keywords: "Робот" ("Robot") or "Тележка" ("Cart"). Functionally, both keywords are recognition activation words, according to the user's preference. The user can also choose the synthetic voice. There are two choices: male and female synthetic voice. The last item on the menu is the choice between online and offline speech recognition. In offline mode, speech recognition is carried out without using the Internet. Activation of offline mode is performed by touching the switch "Offline recognition". Offline speech recognition allows processing continuous speech without an Internet connection, which speeds up the speech recognition process. However, it is worth mentioning that for the offline mode, the user must first download the language pack for the Russian language provided by Google to his/her smartphone, which is a shortened version of the online recognition language pack.

If the command is recognized incorrectly, the user can cancel it by saying "Отмена" ("Cancel") or by touching the corresponding button on the graphical interface.

Below in Figure 7, a general flowchart of the robot's actions is given. After successful processing of the request, the robot starts moving around the store, using a map of the area, and continuous estimation of location based on the Monte Carlo method [55,56] is being performed. After completing a

dialogue interaction flow cycle (user request to goodbye), the robotic cart goes to the base and switches to standby mode.



(**a**) (**b**) (**c**)

Figure 6. Voice interface settings menus: (**a**) The main settings menu; (**b**) The menu of keyword selection; (**c**) The menu of synthesized voice selection.

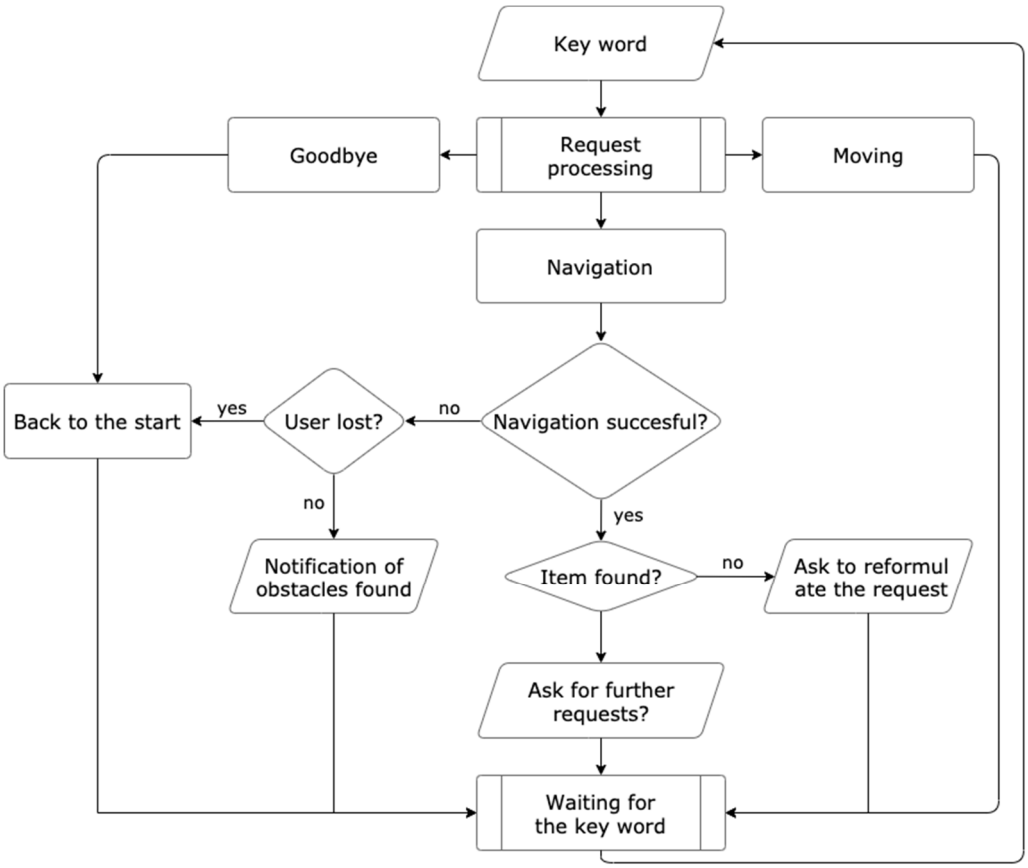


Figure 7. A flowchart presenting user-AMIR interaction scenarios.

The processing of the request itself is carried out by isolating the keywords from the input signal and comparing them with the elements of the dictionary. The robot operates with three dictionaries: products dictionary, departments dictionary, commands dictionary. Each of the products listed in the product dictionary is assigned to each department of the store listed in the department dictionary (see Table 1). The goal of the search algorithm is to determine a specific location that matches the user's request and build an appropriate route.

By continuously determining its position on the map, the AMIR robot builds the most rational route to a point (or a group of points) marked as a particular shop department (e.g., "meat", "dairy products", "baked goods").

The dictionary includes the names of goods without specific brands: "fish", "apples", "eggs", "tea", "juice", etc. One cannot use the names of specific products since such a list could be extremely long. The dictionaries are constructed based on the list of products and departments specified by each store.

4.3. Gesture Interface (Details of Gesture Recognition System Were Previously Published. This Section of the Present Paper Is a Summary of This Work, Briefing the Reader on Key Aspects of It)

The dictionary [57] serves as the main reference point for informants when working on a gesture-based interface. This fundamental work codifies the literary norm for Russian sign language. The use of this edition seems convenient to the authors of this paper because the lexemes included in it are understandable to the overwhelming majority of Russian sign language speakers. At the same time, the subject area "food" does not explicitly refer to either the literary style or to colloquial language or dialects, which guarantees comprehensibility of the gestures even for those speakers who are not familiar with the literary standard of Russian sign language.

The primary list of commands is formed up by exporting text files from Internet navigation menus of a number of local supermarkets. Elaboration of the final vocabulary list is carried out by screening out units containing specific names (brands, manufacturer, ingredients). In addition, the final list does not include products that, according to the personal feelings of the authors of this work, don't enjoy great popularity among customers. Lexical units for which fingerprinting is used are excluded from the vocabulary as well due to the lack of generally accepted gestures. One of the reasons, which has prompted the authors to reduce the final list of gestures is comprehensibility and usability.

4.3.1. Sign Language Synthesis

The sign language synthesis module serves as a gesture output, using an animated 3D avatar. It performs animation of the Russian sign language gestures needed for interaction. After previous experiments with the rule-based sign language synthesis [58] and its implementation in the intelligent information kiosk [59], we have decided on the data-driven synthesis. It allows a higher level of naturalness, which, in the case of hearing-impaired users, ensures also a higher level of intelligibility. To achieve a high-quality synthesis, it is crucial to record a high-quality data set [60], which is possible with the latest motion capture technology. We have taken advantage of the high-quality equipment of our research center.

We have used an optical-based MoCap system consisting of 18 VICON cameras (8xT-20, 4xT-10, 6xVero) for dataset recording and one RGB camera as referential and two Kinects v2 for additional data acquisition. MoCap recording frequency is 120 Hz. The placement of cameras shown in Figure 8 is developed to cover the place in front of the signer in order to avoid occlusions as much as possible and in order to focus on facial expressions. Camera placement is also adjusted for the particular signer to reduce gaps in trajectories caused by occlusions. The layout of the cameras is depicted in Figure 8.

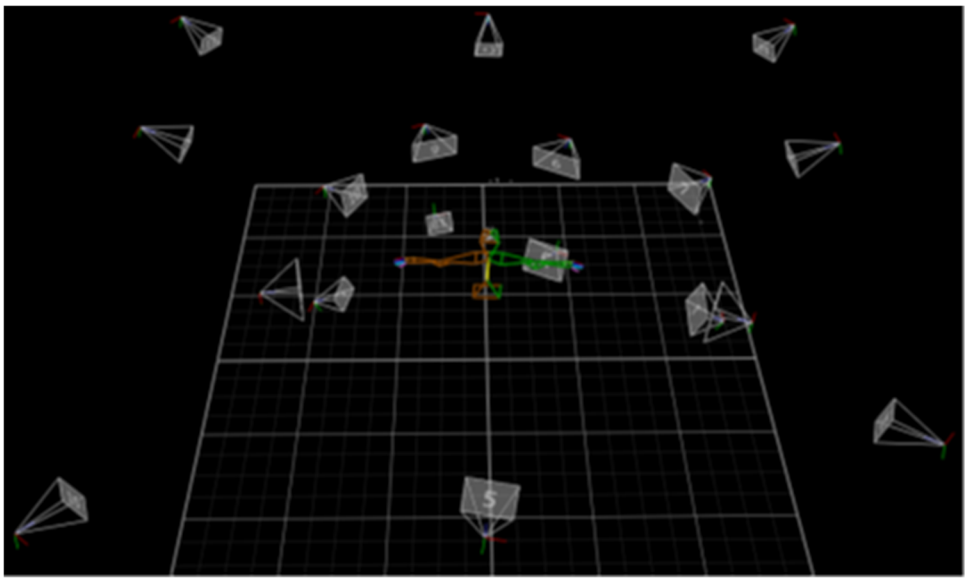


Figure 8. Visualization of MoCap camera layout. View from back and above and the signer is in the middle.

We have recorded approximately 30 min of continuous speech (>200 k frames) and 10 min of dictionary items. All data is recorded by one native sign language expert, who is monitored by another one during the process. The dataset contains 36 weather forecasts. On average, each such forecast is 30 s long and contains 35 glosses. The dictionary contains 318 different glosses. Those dictionary items are single utterances surrounded by the posture with loose hands and arms (rest pose) in order not to be affected by any context.

The markers are placed on the face and fingers. The marker structure is selected to cause minimal disturbance to the signer. We have used different marker sizes and shapes for different body parts. We have tracked the upper body and arms by a pair of markers placed on the axis of joints completed by some referential markers. The positions of markers on the face are selected to follow facial muscles and wrinkles. We have used 8 mm spherical markers around the face, 4 mm hemispherical markers for facial features with the exception of nasolabial folds with 2.5 mm hemispherical markers. Two markers for palm tracking are placed on the index and small finger metacarpals. We have tracked fingers using three 4 mm hemispherical markers per finger placed in the middle of each finger phalanx and thumb metacarpals. The marker setup is depicted in Figure 9a.

Motion capture data are then transferred onto the resulting avatar using the process called retargeting. Data are first translated from a marker structure to the skeleton that is accepted by the animation module. An example of the skeleton structure is depicted in Figure 9.

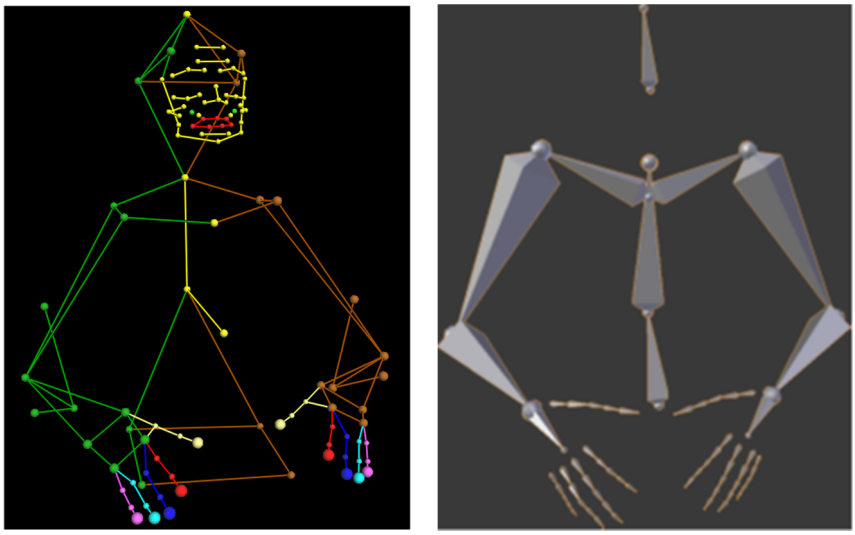
The transitions are synthesized with a constant length, and such an approximation does not correspond with the observed reality. The cubic spline interpolation is also heavily dependent on the annotation's precise selection of the start and the endpoint and also does not respect the nature of the human movement. Examples of a resulting avatar are in Figure 10.

4.3.2. Sign Language Recognition

Using gestures allows contactless interaction with AMIR for various user groups, including people with hearing and vision impairments. A functional diagram of the method of single-hand movements video analysis for recognizing signs of sign language (i.e., isolated commands) is shown in Figure 11.

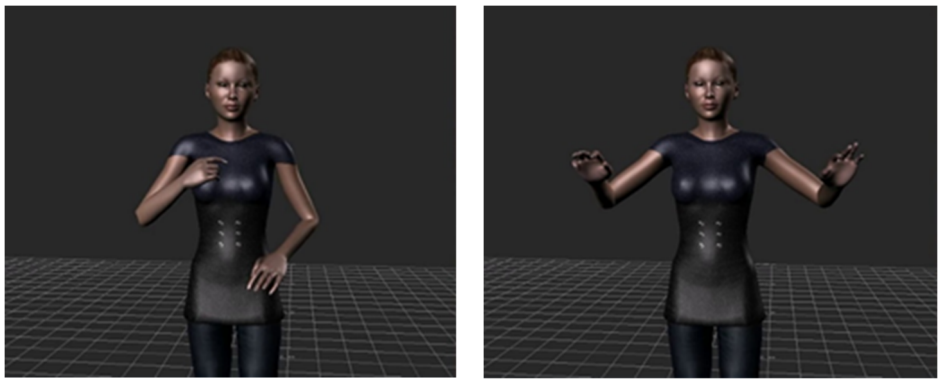
Color video data in MP4 or AVI formats and a depth map in binary format (BIN), as well as text files in the format of the extensible markup language (XML) with 3D and 2D coordinates of skeletal models of signers from the collected and annotated multimedia database (see below, Section 5.1), or color (RGB) video stream and depth map obtained from the Kinect v2 sensor (online mode) are fed to the input

of the developed method in offline mode (testing stage). The method is automatically interrupted if the Kinect v2 sensor is unavailable or the necessary files are not available in the multimedia database; otherwise, cyclic processing of frames is carried out, and a check for a certain frame is performed at each iteration. At this stage, stopping can happen if an error occurs when receiving both RGB video frames and a depth map, as well as if one of the described video streams is stopped.



(**a**) (**b**)

Figure 9. (**a**) Marker setup (data visualization); (**b**) model visualization.



(**a**) (**b**)

Figure 10. Two examples of the signing avatar (**a**,**b**).

The data from the annotated database TheRuSLan are used to train the neural network. Video frames are labeled, and principal handshapes (projections) are used for training (see Section 5.2). Generation of areas containing user images on each 3D frame of the depth map, as well as the calculation of 3D 25-point models of people skeletons, is carried out via a software development kit (SDK) [61,62] of the Kinect sensor, which generates a depth map. Tracking of the nearest user is based on the determination of the nearest 3D skeletal model along the *Z*-axis of the three-dimensional space by calculating the minimum value from all average values of the *Z*-axis of 25-point models of human skeletons. Transformation of a 25-point 3D skeletal model of the nearest user into a 2D 25-point skeletal model is carried out using the Kinect SDK 2.0, which allows you to form 2D regions with the nearest person (see Figure 12). Within the formed rectangular 2D area with the user, a 2D area with his active palm is defined. For this, the MediaPipe model [63] is used.

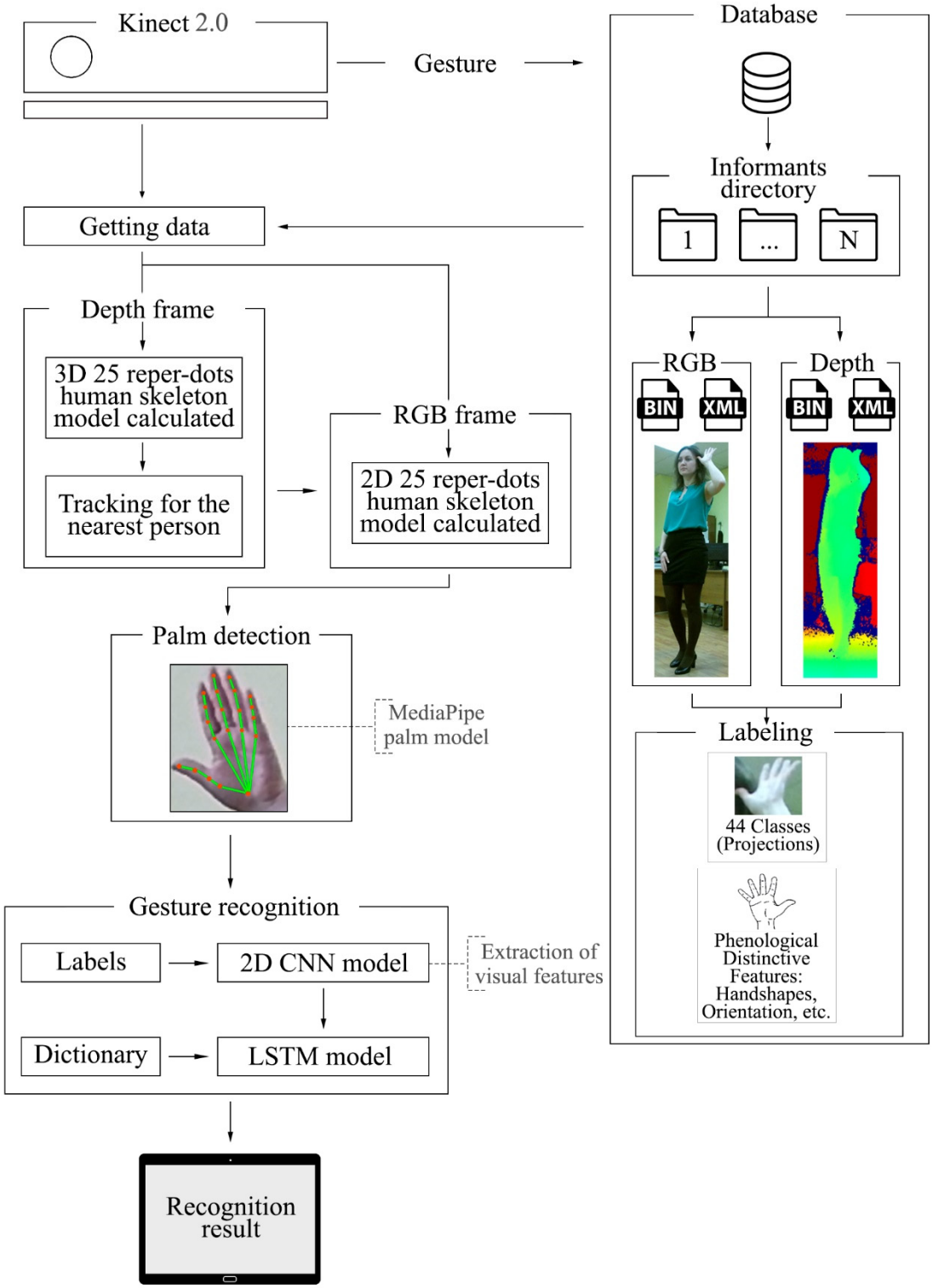


Figure 11. Functional diagram of the sign language recognition method.

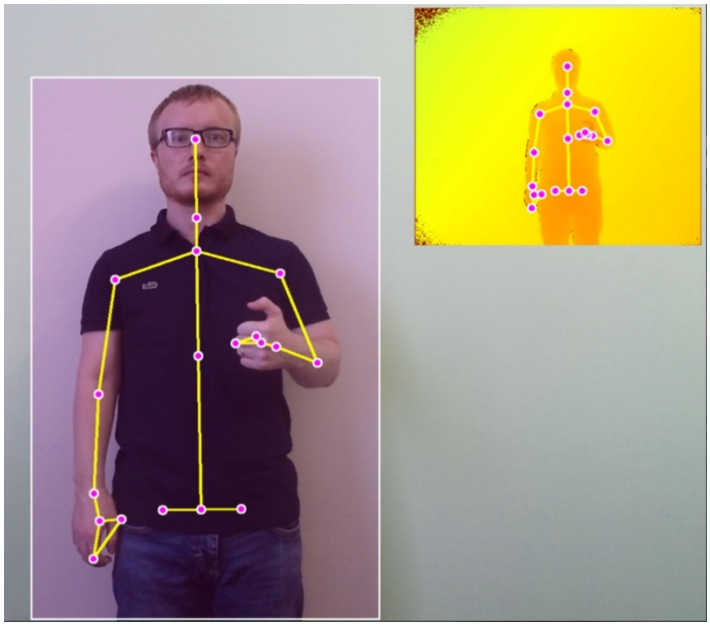


Figure 12. An example of a 2D 25-point human's skeletal model.

In order to extract visual features, a 2D convolutional neural network (2D CNN) is used, with the last fully connected layer of the 2D CNN being ignored for cascade interconnection to a long short-term memory (LSTM) model. The LSTM model is used for gesture recognition. The architecture of the 2D CNN LSTM neural network designed for recognizing individuals' gestures of Russian sign language is presented in Figure 13.

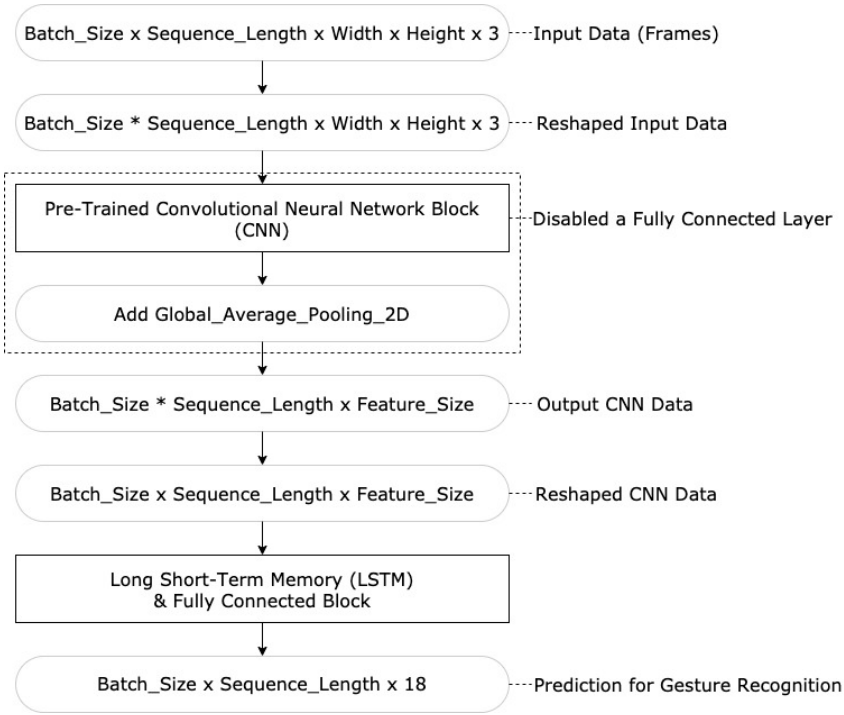


Figure 13. The architecture of the 2D convolutional neural network long short-term memory (2D CNN LSTM) neural network for sign language recognition.

In more detail, the input data for the 2D CNN LSTM neural network are two batch sizes with sequences of the length of 32 frames (64). Each isolated frame from the video sequence has a resolution

that corresponds to the input image size for each separated pre-trained neural network. Next, the input data are resized from 2 × 32 × Width × Height × Channels (3, RGB) to 64 × Width × Height × 3, where Width and Height are the corresponding dimensions of the image. The width and height values are equal and depend on the chosen 2D CNN neural network architecture. Input image resizing is required in order to fit the input size of pre-trained 2D CNN neural network models. All evaluated pre-trained 2D CNN neural network models are initialized with the fully connected layer disabled and the 2D global average pooling added. Each of the 2D CNN models extracts features. Thus, the 2D CNN extracts the features of specific gestures at the output, with the dimension equal to 64 × Feature_Size. Subsequently, the dimensionality of the 2D CNN output is changed from 64×Feature_Size to 2 × 32 × Feature_Size and is fed into the LSTM input of the neural network architecture.

5. Preliminary Experiments and Results

This section presents preliminary experiments conducted during the development of gesture interface development. As repeatedly emphasized by the authors, AMIR may be called a prototype rather than a production-ready robotic platform. This determines the nature of the experiments presented in this section: they are aimed not at the practical application performance of the robotic platform but at testing the key functions of the developed interface. Therefore, it would be correct to talk about preliminary and not full-scale experiments. Preliminary work, i.e., database annotation, is described in Section 5.1, and results are presented in Section 5.2.

5.1. Database Annotation

The main problem with the task of Russian sign language recognition is a lack of resources, such as annotated datasets, corpora, etc. The creation and the annotation of a database is a prerequisite for recognition; thus, a new database of Russian sign language items was created and annotated. A detailed description of the database can be found in the paper [64]. It is worth mentioning that this is the first and the only multimodal database of Russian sign language; all the current electronic collections of Russian sign language items are but mere vocabularies, with the only exception being the corpus of Russian sign language [65], collected at the Novosibirsk State Technical University. That corpus, however, is linguistically oriented and not suitable for machine-learning purposes.

The annotation process is a two-step procedure. At the first stage, all the images are examined, and a set of handshapes and hand positions used by the signers is built up. Currently, there are no researches in which the inventory of Russian sign language handshapes is fully described; thus, seven principal handshapes (hand configuration) with modifications are identified and 11 principal signing areas.

At the second stage, the hand orientation parameter is addressed. The standard HamNoSys classification introduces 18 spatial axes of hands and eight palm orientations [66]. Such instrumentation, being quite a powerful tool, is not appropriate for our purposes, and that's why the annotation procedure is reduced to identifying different projections of handshapes. Handshape projections, as such, are combinations of hand configurations and hand orientation. A total of 44 projections are obtained that can be used for machine learning classification tasks. An example of different projections of two different hand configurations is given in Figure 14 below:

Figure 14. Projections of two handshapes (**a**), (**b**) identified in the collected database.

There is a basic difference between handshapes and hand projections: the former is based on linguistic phonological features (selected fingers and operations with them) and can be used for the linguistic description of Russian sign language, while the latter is based on visual criteria, providing the neural network classification model with as many samples as possible.

5.2. Gesture Recognition Experiments

Various architectures of 2D CNNs combined with different configurations of LSTM are evaluated. All evaluated 2D CNN models tabulated in Table 2 are included in the object recognition module of the Keras open-source library [67]. The number of output clusters of the LSTM model is 512. The dropout of units is performed with 50% probability. Next, a fully connected layer is applied to the number of outputs corresponding to the number of classes, i.e., 18 gesture types. The initial hyperparameters of the learning process are the following: epoch numbers equal to 30, the Adam optimizer with learning rate equal to 0.001. The learning process is stopped when accuracy on the validation set does not increase in three consecutive epochs.

Table 2. Gesture recognition accuracy (%) for different 2D CNN + LSTM models.

Input Image Size Width × Height × Channels	Feature Size	Model 2D CNN-LSTM	Accuracy (%)
299 × 299 × 3	2048	Xception	80.03
224 × 224 × 3	512	VGG16	72.58
224 × 224 × 3	512	VGG19	73.19
224 × 224 × 3	2048	ResNet152V2	76.11
299 × 299 × 3	2048	InceptionV3	75.92
299 × 299 × 3	1536	InceptionResNetV2	81.44
224 × 224 × 3	1280	MobileNetV2	72.47
224 × 224 × 3	1664	DenseNet169	76.54
331 × 331 × 3	4032	NASNetLarge	84.44
224 × 224 × 3	1280	EfficientNetB0	70.32
528 × 528 × 3	2559	EfficientNetB7	87.01

CNN—convolutional neural network, LSTM—long short-term memory neural network; VGG—a particular convolutional neural network architecture (stands for "Visual Geometry Group").

Transfer learning is performed using labeled data (see MediaPipe [63]) with hand shapes from the TheRuSLan database [64] (54 gestures). The dataset is split into training and test samples in an approximate train/test ratio of 80%/20%. Half of the test data are used as validation data. The gesture recognition accuracy results for different pre-trained CNN models that are tabulated in Table 2. The best accuracy is shown in bold. The results in Table 2 are listed in increasing order of the size of the input image.

As shown in Table 2, the best gesture recognition accuracy, 87.01%, is achieved by the 2D CNN EfficientNetB7 model [68]. Experiments have shown that the accuracy of hand gesture recognition is influenced by the depth of the 2D CNN layers and the size of extracted features.

The process of sending a control command corresponding to the recognized gesture or a voice command to AMIR is made via Wi-Fi 802.11 connection under TCP/IP protocol and a dedicated port. AMIR's corresponding action is based on the result of recognized audio or video information from the interacting user. An example of Russian sign language gesture recognition can be found on YouTube video hosting (https://youtu.be/dWbrV7eVqn4).

The results obtained in these tests are conducted under laboratory conditions, and further modifications of the architecture of the proposed model as well as the collection of additional data will assist in achieving the even higher performance of the system. In addition, the design of the proposed gesture recognition method allows the replacement of the Kinect v2 sensor with more modern versions, such as the Azure Kinect.

6. Conclusions

This article presented a multimodal user interface for interacting with the AMIR robotic cart. AMIR is an assistive robotic platform with the architecture of the interface being focused on the needs of elders and/or people with hearing impairments. Among the strong points of the AMIR project are enhanced marketability: the robotic carts overview (Section 2) demonstrates the high demand for development of this kind, and lightweight technical solutions: all the modules of the device are easy to assemble.

AMIR is the only project that deals with real-time Russian sign language recognition/synthesis and can be effectively used as a service robot, serving the needs of a wide range of supermarket customers. It should be emphasized that although the vocabulary of gestures used in AMIR to interact with deaf users is very limited, its potential size is limited only by the collected database. The dictionary is based on TheRuSLan database, which is actively expanding at the moment; in theory, the AMIR project's gesture interface is quite capable of capturing most of the subject vocabulary on the topic of "goods in the supermarket". TheRuSLan database is currently one of the few resources on Russian sign language, which is suitable for the training of Russian sign language recognition systems and allowing linguistic research.

Among the main problems faced by the authors of this study in collecting the database of gestures are the following:

(a) the complexity of recording gestures on a subject topic: there are no databases on the selected subject topic as such, and the existing multimedia dictionaries of Russian sign languages either do not include a number of lexemes relevant to the AMIR project or these lexemes are represented by variants characteristic of the non-Petersburg dialect area;

(b) the complexity of machine learning for sign languages: typical problems are the variability of the same gesture, the relatively small size of the hands compared to the human torso, numerous occlusions, noises, and differences in the tones of the speakers' skin.

The first problem can be solved solely by collecting new databases. AMIR developers have a number of publications devoted to this topic, and in the future, it is planned to create a full-fledged corpus of Russian sign language. As for the second problem, it seems to the authors of this article that the developed Deep neural network architecture quite successfully copes with the task (see Section 5).

There is a possibility of using such an interface for gestural communication with people who do not suffer from hearing impairments. Indeed, non-verbal communication (gestures, body language) is an integral part of everyday communication. At the same time, in some cases, the gesture is more preferable than the verbal command. In general, the user seeks to select the information transmission channel that best suits the communicative context. Besides, if we are talking exclusively about the gesture modality, then the use of gestures should not be a forced measure, but a natural, unobtrusively complementing other modalities—or preferred in a specific communicative situation—way of exchanging information.

It should be noted that the assistive characteristics of the AMIR robot are not limited to the use of the Russian sign language. Freeing the user from pushing the cart in front of him or carrying a full basket significantly reduces shopping fatigue, and the user-requested navigation system reduces the time spent in the store. Such a robotic platform provides support not only for the hearing impaired but also for the elderly or people with various diseases that complicate long-term movements around the store with loads in the form of a basket or cart. The prospects for the implementation of AMIR are thus in no way exclusively related to the categories of persons with hearing impairments. It seems that such a robotic platform will be relevant in any supermarket.

Author Contributions: A.A. prepared the related work analysis on the topic of assistive and mobile robotics (Section 2). D.R. and I.K. (Irina Kipyatkova) were responsible for the general description of the interface, gesture, and voice recognition approach. N.P., together with A.S., contributed to the article by providing all the technical details of the AMIR prototype (Section 3). I.K. (Ildar Kagirov) worked on Conclusions, Introduction, and the description of the gesture vocabulary (Section 4). A.K., M.Z., and I.M. contributed to the writing of the paper and conducted the general supervision of the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financially supported by the Ministry of Science and Higher Education of the Russian Federation, the agreement No. 14.616.21.0095 (ID Number: RFMEFI61618X0095), as well as a part of the Russian state research No. 0073-2019-0005.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wada, K.; Shibata, T. Living with seal robots—Its sociopsychological and physiological influences on the elderly at a care house. *IEEE Trans. Robot.* **2007**, *23*, 972–980. [CrossRef]
2. Wada, K.; Shibata, T.; Musha, T.; Kimura, S. Robot therapy for elders affected by dementia. *IEEE Eng. Med. Biol. Mag.* **2008**, *27*, 53–60. [CrossRef]
3. Wada, K.; Shibata, T. Social and physiological influences of robot therapy in a care house. *Interact. Stud.* **2008**, *9*, 258–276. [CrossRef]
4. Bemelmans, R.; Gelderblom, G.J.; Jonker, P.; Witte, L. The potential of socially assistive robotics in care for elderly, a systematic review. *Hum. Robot Pers. Relatsh.* **2011**, *59*, 83–89.
5. Bemelmans, R.; Gelderblom, G.J.; Jonker, P.; Witte, L. Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *J. Am. Med Dir. Assoc.* **2012**, *13*, 114–120. [CrossRef] [PubMed]
6. Maldonado-Bascon, S.; Iglesias-Iglesias, C.; Martín-Martín, P.; Lafuente-Arroyo, S. Fallen people detection capabilities using assistive robot. *Electronics* **2019**, *8*, 915. [CrossRef]
7. Lotfi, A.; Langensiepen, C.; Yahaya, S.W. Socially Assistive Robotics: Robot Exercise Trainer for Older Adults. *Technologies* **2018**, *6*, 32. [CrossRef]
8. Piezzo, C.; Suzuki, K. Feasibility Study of a Socially Assistive Humanoid Robot for Guiding Elderly Individuals during Walking. *Future Internet* **2017**, *9*, 30. [CrossRef]
9. Sumiya, T.; Matsubara, Y.; Nakano, M.; Sugaya, M. A mobile robot for fall detection for elderly-care. *Procedia Comput. Sci.* **2015**, *60*, 870–880. [CrossRef]
10. Antonopoulos, C.P.; Kelverkloglou, P.; Voros, N.; Fotiou, I. Developing Autonomous Cart Equipment and Associated Services for Supporting People with Moving Disabilities in Supermarkets: The EQUAL Approach. In Proceedings of the Project Track of IISA 2019, Patras, Greece, 15–17 July 2019; pp. 20–24.
11. Broekens, J.; Heerink, M.; Rosendal, H. Assistive social robots in elderly care: A review. *Gerontechnology* **2009**, *8*, 94–103. [CrossRef]
12. Feil-Seifer, D.; Mataric, M.J. Defining socially assistive robotics. In Proceedings of the IEEE 9th International Conference on Rehabilitation Robotics, Chicago, IL, USA, 28 June–1 July 2005.
13. Ryumin, D.; Kagirov, I.; Železný, M. Gesture-Based Intelligent User Interface for Control of an Assistive Mobile Information Robot. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12336 LNAI, pp. 126–134.

14. Ryumin, D.; Kagirov, I.; Ivanko, D.; Axyonov, A.; Karpov, A. Automatic detection and recognition of 3D manual gestures for human-machine interaction. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2019**, *XLII-2/W12*, 179–183. [CrossRef]
15. Matsuhira, N.; Ozaki, F.; Tokura, S.; Sonoura, T.; Tasaki, T.; Ogawa, H.; Sano, M.; Numata, A.; Hashimoto, N.; Komoriya, K. Development of robotic transportation system-Shopping support system collaborating with environmental cameras and mobile robots. In Proceedings of the 41st International Symposium on robotics and 6th German Conference on Robotics (ROBOTIK 2010), Munich, Germany, 7–9 June 2010; pp. 1–6.
16. Chiang, H.; Chen, Y.; Wu, C.; Kau, L. Shopping assistance and information providing integrated in a robotic shopping cart. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE-TW), Taipei, Taiwan, 12–14 June 2017; pp. 267–268.
17. Raulcezar, A.; Linhares, B.A.; Souza, J.R. Autonomous Shopping Cart: A New Concept of Service Robot for Assisting Customers. In *2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE)*; IEEE: Piscataway, NJ, USA, 2018.
18. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **1994**, *4*, 65–85. [CrossRef]
19. Davendra, D. (Ed.) *Traveling Salesman Problem: Theory and Applications*; InTech: Munich, Germany, 2010. Available online: https://www.intechopen.com/books/traveling-salesman-problem-theory-and-applications (accessed on 10 September 2020).
20. Su, H.; Zhang, Y.; Li, J.; Hu, J. The shopping assistant robot design based on ROS and deep learning. In Proceedings of the 2nd International Conf. Cloud Computing and Internet of Things, Dalian, China, 22–23 October 2016; pp. 173–176.
21. Kobayashi, Y.; Yamazaki, S.; Takahashi, H.; Fukuda, H.; Kuno, Y. Robotic shopping trolley for supporting the elderly. In Proceedings of the AHFE 2018 International Conference on Human Factors and Ergonomics in Healthcare and Medical Devices, Loews Sapphire Falls Resort at Universal Studios, Orlando, FL, USA, 21–25 July 2018; pp. 344–353.
22. Garcia-Arroyo, M.; Marin-Urias, L.F.; Marin-Hernandez, A.; Hoyos-Rivera, G.D.J. Design, integration, and test of a shopping assistance robot system. In Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, USA, 5–8 March 2012; pp. 135–136.
23. Marin-Hernandez, A.; de Jesus Hoyos-Rivera, G.; Garcia-Arroyo, M.; Marin-Urias, L.F. Conception and implementation of a supermarket shopping assistant system. In Proceedings of the 11th Mexican International Conference on Artificial Intelligence, San Luis Potosi, Mexico, 27 October–4 November 2012; pp. 26–31.
24. Sales, J.; Marti, J.V.; Marín, R.; Cervera, E.; Sanz, P.J. CompaRob: The shopping cart assistance robot. *Int. J. Distrib. Sens. Netw.* **2016**, *12*, 16. [CrossRef]
25. Ikeda, H.; Kawabe, T.; Wada, R.; Sato, K. Step-Climbing Tactics Using a Mobile Robot Pushing a Hand Cart. *Appl. Sci.* **2018**, *8*, 2114. [CrossRef]
26. Elgendy, M.; Sik-Lanyi, C.; Kelemen, A. Making Shopping Easy for People with Visual Impairment Using Mobile Assistive Technologies. *Appl. Sci.* **2019**, *9*, 1061. [CrossRef]
27. Andò, B.; Baglio, S.; Marletta, V.; Crispino, R.; Pistorio, A. A Measurement Strategy to Assess the Optimal Design of an RFID-Based Navigation Aid. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 2356–2362. [CrossRef]
28. Kesh, S. Shopping by Blind People: Detection of Interactions in Ambient Assisted Living Environments using RFID. *Int. J. Wirel. Commun. Netw. Technol.* **2017**, *6*, 7–11.
29. Kulyukin, V.; Gharpure, C.; Nicholson, J. Robocart: Toward robot-assisted navigation of grocery stores by the visually impaired. In Proceedings of the International Conference on Intelligent Robots and Systems (IROS-2005), Edmonton, AB, Canada, 2–6 August 2005; pp. 2845–2850.
30. Kuipers, B. The spatial semantic hierarchy. *Artif. Intell.* **2000**, *119*, 191–233. [CrossRef]
31. Azenkot, S.; Zhao, Y. Designing smartglasses applications for people with low vision. *ACM SIGACCESS Access. Comput.* **2017**, *119*, 19–24. [CrossRef]
32. López-de-Ipiña, D.; Lorido, T.; López, U. Indoor navigation and product recognition for blind people assisted shopping. In Proceedings of the International Workshop on Ambient Assisted Living (IWAAL), Torremolinos-Málaga, Spain, 8–10 June 2011; pp. 33–40.
33. Tomizawa, T.; Ohya, A.; Yuta, S. Remote shopping robot system: Development of a hand mechanism for grasping fresh foods in a supermarket. In Proceedings of the Intelligent Robots and Systems (IROS-2006), Bejing, China, 3–8 November 2006; pp. 4953–4958.

34. Tomizawa, T.; Ohba, K.; Ohya, A.; Yuta, S. Remote food shopping robot system in a supermarket-realization of the shopping task from remote places. In Proceedings of the International Conference on Mechatronics and Automation (ICMA-2007), Harbin, Heilongjiang, China, 5–8 August 2007; pp. 1771–1776.
35. Wang, Y.C.; Yang, C.C. 3S-trolley: A lightweight, interactive sensor-based trolley for smart shopping in supermarkets. *IEEE Sens. J.* **2016**, *16*, 6774–6781. [CrossRef]
36. Hart, P.E.; Nilsson, N.J.; Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 100–107. [CrossRef]
37. Kumar, A.; Gupta, A.; Balamurugan, S.; Balaji, S.; Marimuthu, R. Smart shopping cart. In Proceedings of the 2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS 2017), VIT University, Vellore, India, 10–12 August 2017; pp. 1–4.
38. Gangwal, U.; Roy, S.; Bapat, J. Smart shopping cart for automated billing purpose using wireless sensor networks. In Proceedings of the Seventh International Conference on Sensor Technologies and Applications (SENSORCOMM 2013), Barcelona, Spain, 25–31 August 2013; pp. 168–172.
39. Athauda, T.; Marin, J.C.L.; Lee, J.; Karmakar, N.C. Robust low-cost passive UHF RFID based smart shopping trolley. *IEEE J. Radio Freq. Identif.* **2018**, *2*, 134–143. [CrossRef]
40. Kanda, T.; Shiomi, M.; Miyashita, Z.; Ishiguro, H.; Hagita, N. A communication robot in a shopping mall. *IEEE Trans. Robot.* **2010**, *26*, 897–913. [CrossRef]
41. Kanda, T.; Shiomi, M.; Miyashita, Z.; Ishiguro, H.; Hagita, N. An affective guide robot in a shopping mall. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (HRI-09), La Jolla, CA, USA, 11–13 March 2009; pp. 173–180.
42. Shiomi, M.; Kanda, T.; Glas, D.F.; Satake, S.; Ishiguro, H.; Hagita, N. Field trial of networked social robots in a shopping mall. In Proceedings of the Intelligent Robots and Systems (IROS-2009), Hyatt Regency St. Louis Riverfront, St. Louis, MO, USA, 11–15 October 2009; pp. 2846–2853.
43. Chen, C.C.; Huang, T.-C.; Park, J.H.; Tseng, H.-H.; Yen, N.Y. A smart assistant toward product-awareness shopping. *Pers. Ubiquitous Comput.* **2014**, *18*, 339–349. [CrossRef]
44. Surdofon: Russian Sign Language On-Line Translation. Available online: https://surdofon.rf (accessed on 14 August 2020).
45. Falconer, J. Humanoid Robot Demonstrates Sign Language. Available online: https://spectrum.ieee.org/automaton/robotics/humanoids/ntu-taiwan-humanoid-sign-language (accessed on 10 August 2020).
46. Kose, H.; Yorganci, R. Tale of a robot: Humanoid robot assisted sign language tutoring. In Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots, Bled, Slovenia, 26–28 October 2011; pp. 105–111.
47. Hoshino, K.; Kawabuchi, I. A humanoid robotic hand performing the sign language motions. In Proceedings of the 2003 International Symposium on Micromechatronics and Human Science (MHS-2003), Tsukuba, Japan, 20 October 2003; pp. 89–94.
48. Koubaa, A. (Ed.) *Robot Operating System (ROS): The Complete Reference*; Springer: Cham, Switzerland, 2019; Volume 4.
49. Montemerlo, M.; Thrun, S.; Koller, D.; Wegbreit, B. FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. In Proceedings of the Eighteenth National Conference on Artificial Intelligence, Edmonton, AB, Canada, 28 July–1 August 2002; pp. 593–598.
50. Montemerlo, M.; Koller, S.T.D.; Wegbreit, B. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In Proceedings of the Int. Conf. on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 9–15 August 2003; pp. 1151–1156.
51. Grisetti, G.; Stachniss, C.; Burgard, W. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Trans. Robot.* **2007**, *23*, 34–46. [CrossRef]
52. Grisetti, G.; Tipaldi, G.D.; Stachniss, C.; Burgard, W.; Nardi, D. Fast and accurate SLAM with Rao–Blackwellized particle filters. *Robot. Auton. Syst.* **2007**, *55*, 30–38. [CrossRef]
53. Brock, O.; Khatib, O. High-speed navigation using the global dynamic window approach. In Proceedings of the 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C), Detroit, MI, USA, 10–15 May 1999; pp. 341–346.
54. Documentation for Android Developers, SpeechRecognizer. Available online: https://developer.android.com/reference/android/speech/SpeechRecognizer (accessed on 23 July 2020).
55. Karger, D.R.; Stein, C. A New Approach to the Minimum Cut Problem. *J. ACM* **1996**, *43*, 601–640. [CrossRef]

56. Kudelic, R. Monte-Carlo randomized algorithm for minimal feedback arc set problem. *Appl. Soft Comput.* **2016**, *41*, 235–246. [CrossRef]
57. Geilman, I.F. *Russian Sign Language Dictionary*; Prana: St. Petersburg, Russia, 2004. (In Russian)
58. Kanis, J.; Zahradil, J.; Jurčíček, F.; Müller, L. Czech-sign speech corpus for semantic based machine translation. In Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006), Brno, Czech Republic, 11–15 September 2006; pp. 613–620.
59. Aran, O.; Campr, P.; Hrúz, M.; Karpov, A.; Santemiz, P.; Zelezny, M. Sign-language-enabled information kiosk. In Proceedings of the 4th Summer Workshop on Multimodal Interfaces eNTERFACE, Orsay, France, 4–29 August 2008; pp. 24–33.
60. Jedlička, P.; Krňoul, Z.; Kanis, J.; Železný, M. Sign Language Motion Capture Dataset for Data-driven Synthesis. In Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (LREC-2020), Marceille, France, 11–16 May 2020; pp. 101–106.
61. Kinect for Windows. Available online: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn782037(v=ieb.10) (accessed on 24 July 2020).
62. Rahman, M. *Beginning Microsoft Kinect for Windows SDK 2.0: Motion and Depth Sensing for Natural User Interfaces*; Apress: Berkeley, CA, USA, 2017; pp. 41–76.
63. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172.
64. Kagirov, I.; Ivanko, D.; Ryumin, D.; Axyonov, A.; Karpov, A. TheRuSLan: Database of Russian Sign Language. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC-2020), Marceille, France, 11–16 May 2020; pp. 6079–6085.
65. Russian Sign Language Corpus. Available online: http://rsl.nstu.ru/site/index/language/en (accessed on 7 December 2020).
66. Prillwitz, S.; Leven, R.; Zienert, H.; Hanke, T.; Henning, J. *HamNoSys*; Version 2.0; Hamburg Notation System for Sign Languages; An Introductory Guide; Signum Verlag: Hamburg, Germany, 1989.
67. Keras Applications. Available online: https://keras.io/api/applications (accessed on 24 July 2020).
68. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.