Development of a Verbal Robot Hand Gesture Recognition System

CHINGIS KENSHIMOV, TALGAT SUNDETOV, MURAT KUNELBAYEV,
ZHAZIRA AMIRGALIYEVA, DIDAR YEDILKHAN, OMIRLAN AUELBEKOV

Institute of Information and Computing Technologies CS MES RK, Almaty, KAZAKHSTAN.

Abstract: - This article analyzes the most famous sign languages, the correlation of sign languages, and also considers the development of a verbal robot hand gesture recognition system in relation to the Kazakh language. The proposed system contains a touch sensor, in which the contact of the electrical property of the user's skin is measured, which provides more accurate information for simulating and indicating the gestures of the robot hand. Within the framework of the system, the speed and accuracy of recognition of each gesture of the verbal robot are calculated. The average recognition accuracy was over 98%. The detection time was 3ms on a 1.9 GHz Jetson Nano processor, which is enough to create a robot showing natural language gestures. A complete fingerprint of the Kazakh sign language for a verbal robot is also proposed. To improve the quality of gesture recognition, a machine learning method was used. The operability of the developed technique for recognizing gestures by a verbal robot was tested, and on the basis of computational experiments, the effectiveness of algorithms and software for responding to a verbal robot to a voice command was evaluated based on automatic recognition of a multilingual human voice. Thus, we can assume that the authors have proposed an intelligent verbal complex implemented in Python with the CMUSphinx communication module and the PyOpenGL graphical command execution simulator. Robot manipulation module based on 3D modeling from ABB.

Key-Words: - Verbal robot, accurate recognition, hand gesture, Kinect.
Received: March 15, 2021. Revised: October 1, 2021. Accepted: October 25, 2021. Published: November 11, 2021.

1 Introduction

In the field of computer vision and machine learning, hand gesture recognition for human-computer interaction is an area of active research [1]. Identifying specific gestures and using them to convey information or control a device is one of the main studies in gesture recognition. Gestures should be modeled in spatial and temporal areas, where the posture of the hand is the static structure of the hand, and the gesture is the dynamic movement of the hand. One of the most important means of communication in everyday human life, as well as with the continuous development of image and video processing technologies, studies of human-machine interaction through gesture recognition have led to the use of such technology in a very wide range of possible applications [2,3].

In work [4,5], virtual reality was developed and studied, which allows realistically manipulating virtual objects with the help of hands. The article [6] developed gestures used to interact with robots and control robots, in which gestures can control the robot's hand and arm movements to reach and manipulate real objects, as well as its movement around the world.

The work [7,8,9] has developed an application for desktop and tablet PCs in which gestures can provide an alternative interaction between the mouse and keyboard. Many gestures for desktop computing tasks include manipulating graphics or annotating and editing documents using pen-based gestures.

The article[10] has developed games that track the position of the player's hand or body to control the movement and orientation of interactive game objects, as well as the use of gestures to control the movement of avatars in the virtual world. Microsoft introduced Kinect [11], which is able to track the user's entire body to control games.

The work [12,13,14,15] has developed sign languages that are highly structured; they are very suitable as testing grounds for vision-based algorithms.

Sign language, for example, is the most natural way of communicating among deaf people, although it has been observed that they have difficulty interacting with normal people. Sign language consists of a dictionary of signs in the same way that spoken language consists of a dictionary of words. Sign languages are not standard and universal and grammars differ from country to country. The article [16] developed Portuguese Sign Language (PSL), which includes hand movements, body movements and facial

expressions. The goal of the developed sign language recognition system is to provide an efficient and accurate way of converting sign language into text or voice. And also a manual has been developed for hearing-impaired children to interact with computers (sign language recognition). Vijay et al. [17] developed hand gesture communication and divided it into two types of approaches: vision based approaches and data glove techniques. This development of the article focused on creating a vision-based approach to implement a system capable of performing posture and gesture recognition for real-time applications. In their studies [18,219], the researchers used visual input in context, which allows remote communication with a computer without the intervention of physical contact or any additional devices. Hasanuzzaman et al., [20] have developed efficient real-time gesture recognition systems to perform human-like interfaces between humans and robots.

In the context of globalization, the task of creating an international standardization platform for standardizing the sign languages of different countries and the family of sign languages will become more urgent than ever. In different countries, their states, as well as internationally, are taking various measures to support people with disabilities. So in 1951, the World Federation of the Deaf (WFH, World Federation of the Deaf) appeared, as a result of which the participants of the first World Congress of the Deaf decided to standardize sign languages in order to ensure equal communication rights for all participants speaking different languages, especially for people from the number of deaf people at international conferences, symposia and other events for all participants.

At the initiative of the World Federation of the Deaf, a common international sign language was developed, where, in order to create a nucleus of similar gestures, experts analyzed and studied the features of the most popular sign languages. As a result, the first International Sign Language (gestuno) dictionary was published in 1965 and contained 300 signs; in 1973, the World Federation of the Deaf released a simplified sign language dictionary; The 3rd edition of 1975 already included 1,500 gestures.

Naturally, the program of mastering, development and improvement rigidly presupposes understanding and effort on the part of representatives of different sign languages. This language is now used at European international meetings and conferences for the deaf.

It should be noted that several problems arose with the use of stiffness: none of the published dictionaries described the grammatical basis of the system; the artificial principles of the formation of new vocabulary are not explained; the vocabulary of the dictionary was completely based on four sign languages - British, Italian, American and Russian; there were no gestures from Asian, African and South American national sign languages.

In the process of creating an international sign language, an important role is played by the creation of a list of basic words (list of swords), which can be interpreted for different sign languages with similar images. Thus, the Swadesh list (proposed by the American linguist M. Swadesh) is a tool used to assess the degree of kinship between different languages based on the similarity of the basic vocabulary. The minimum set of essential vocabulary is contained in the 100-word Swadesh list.

Currently, scientists and practitioners are conducting research on the Kazakh sign language, at the same time, work is intensifying on the use of robots to represent the Kazakh sign language.

When it comes to an international platform for the harmonization of different sign languages based on a dedicated core of gestures from different languages, it is necessary to note the influence of such popular families of sign languages. This is primarily the family of the French sign language, which includes several sign languages, including Dutch, Irish, Russian, Brazilian sign languages. The British Sign Language family (Australian, New Zealand and British) is very similar to American Sign Language, at about 30%. The Japanese Sign Language family includes Japanese and Taiwanese Sign Language.

Isolated sign languages are also known, such as Albanian, Armenian, Afghan Vietnamese, Mongolian, Nicaraguan, Mali, Hawaiian, etc. sign languages. The Internet resource www.intersignuniversity.com shows Neapali, Israeli, South Korean and Indian sign languages. Below are the alphabets of the most famous sign languages.

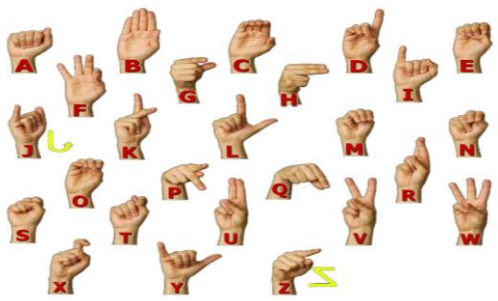

Figure 1: American Sign Alphabet

Figure 2: British Sign Alphabet

The correlation of the similarity of sign languages can be seen in the following table. Kazakh and Russian sign languages are the most similar, apparently this was influenced by the joint existence of the two countries for more than 70 years within the framework of one country - the Soviet Union.

Table 1. Correlation matrix

Sign languages	Kazakh	Russian	Ameri-can	French	British	Australian
Kazakh	100%	97%	61%	54%	21%	17%
Russian	97%	100 %	64%	58%	23%	21%
American	61%	64%	100%	73%	34%	28%
French	54%	58%	73%	100%	46%	40%
British	21%	23%	34%	46%	100%	82%
Austra-lian	17%	21%	28%	40%	82%	100%

2 Research methodology. Hand gesture recognition using the kinect sensor

2.1 Hardware design of an automated verbal robot based on the "Inmoov" platform

Figure 3a, b shows a diagram of the developed 3D printed inexpensive verbal robot. The height of the robot is approximately 170 cm, which is similar to the normal height of an adult. This robot consists of two parts: the 3D printed front of the verbal robot, the 3D printed back of the verbal robot and the mobile base. The mobile base is made of iron and is driven by three 750 W motors. For a life-size verbal robot with strong mobility, this is very cost effective.

As shown in Table 2, the verbal robot has 50 degrees of latitude, 27 motors, 25 servos with different loads and 3 center motors, specially designed for the electrical control of the 24V system, a 16Ah lithium battery pack is installed in the mobile base, and provides all the electrical operation of the robot.

Table 2. Degrees of freedom of verbal robot

Part	Degree of freedom	Motors
Right and left hand	30	10
Wrist	2	2
Elbow	2	2
Shoulder	6	6
Head	3	3
Waist	1	2
Mobile base	6	3

The advantages of the wheeled robot: they are faster, more stable, easy manageable, more efficient and can provide with more payload in implementation, more degree of freedom allows to move efficiently diagonally, right, left and forward, backward.

The verbal robot consists of 1-head, 2- shoulder, 3- face recognition device, 4- elbow, 5- torso, 6-hand, 7- arms, 8- platform for controlling the entire system, 9- omnidirectional mobile platform, 10-touch screen.

The mobile platform contains three universal wheels, each wheel spaced 120 degrees apart to make the robot static and with three distinctive motor wheels to make the robot move.

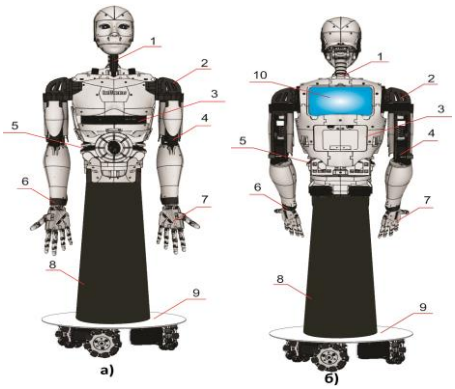

Figure 3: a) 3D printed front side of the verbal robot; b) 3D printed back side of the verbal robot.

2.2 Hardware design of an automated verbal robot based on the "Inmoov" platform

This software is designed to be modular in order to match the hardware. The modular drive is easy to operate and maintain. Six modules have six control panels, and all of these six modules are linked and used in RS485 to communicate with the main Arduino Mega 2560.

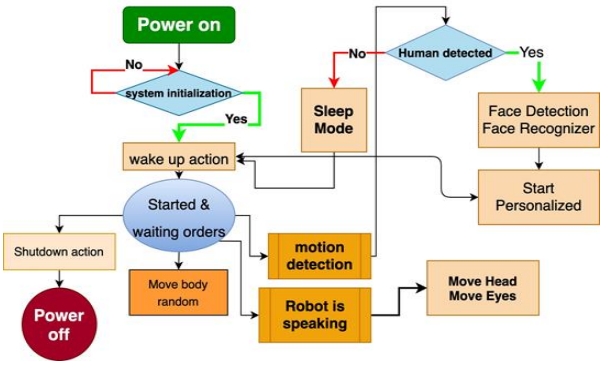

Figure 4:Verbal robot control algorithm

As we can see from Figure 4, when the robot is turned on, the system begins to initialize. After the whole system is connected to the robot modules with 50 degrees of freedom, it starts to move in a chaotic manner, and in parallel, the platform for controlling the entire system will have voice guidance, and in parallel it will read the motion sensor data. If there is movement, then the face recognition device recognizes the moving face of the person. Further, the personalization of the person begins and will conduct a dialogue with the person and a database is created for each person with whom the robot communicated. After the dialogue is over, the data is sent to the propulsion system. If the sensor does not recognize movement, then the robot goes into sleep mode. After sleep mode at a certain time, the entire system is turned off.

Figure 5 shows the proposed structure of the verbal robot software architecture. The software is divided into five layers, the user layer, the software module layer, the data processing layer, the communication layer, and the Executive layer. The user layer includes a function that directly interacts with the robot operator, touch screen, Kinect sensor, and web monitoring system.

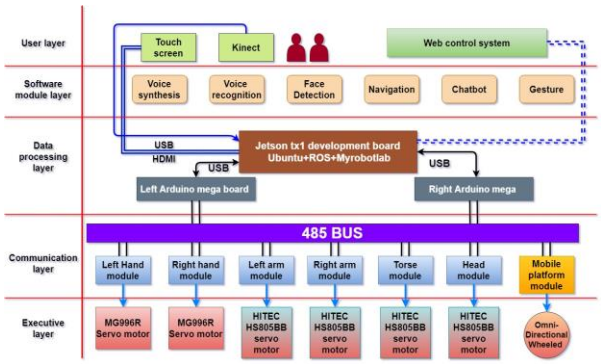

Figure 5: Verbal robot software architecture "INMOOV"

The software module layer consists of software modules for speech recognition, speech synthesis, face recognition, navigation modules and a chat bot that simulates a real conversation with the user, as well as a software module for gestures. There is a logical connection and intelligent control between them. In “INMOOV,” the humanoid robot Jetson-tx1, running Ubuntu and ROS act as the main controller and coordinator. Based on the rich ROS libraries, it becomes easier to get the human skeleton and other sensor information from the Kinect, camera, and microphones. The data processing layer consists of Jetson-tx1, which can receive commands from Kinect. The communication layer consists of Arduino Mega 2560 and six nested dispatcher modules. The executive layer consists of servo motors and 120 degree three motors. The RS485 has a simple structure and can have many slave modules. In the RS485 network Arduino Mega, as the main computer and control panel of each module, serves as a computer. The main computer is responsible for the control command, for acquiring data and for executing the control signal. When the control panels communicate with the host computer, it follows the Modbus protocol, which uses the mechanism to eliminate communication errors. In this network, each control panel serves as a node, which has a different ID. ID can be configured using the onboard Dip Switch.

3 Robot hand gesture recognition system

Hand gesture recognition is an important research issue in the field of human-machine interaction, due to its widespread use in virtual reality, language gesture recognition and computer

games. Despite the large amount of previous work, the creation of an accurate hand gesture recognition system that is applicable to real-world applications remains a difficult problem.

The hand gesture recognition challenge addresses two complex problems: hand detection and gesture recognition, namely, how to reliably detect the hand and how to effectively and accurately identify the hand gesture.

The novelty of this research is that the recognition of hand gestures comprises at least one or more touch sensors for measuring the contact of the electrical property of the user's skin, based on which different hand gestures are recognize Figure 6 shows a hand with touch sensors.

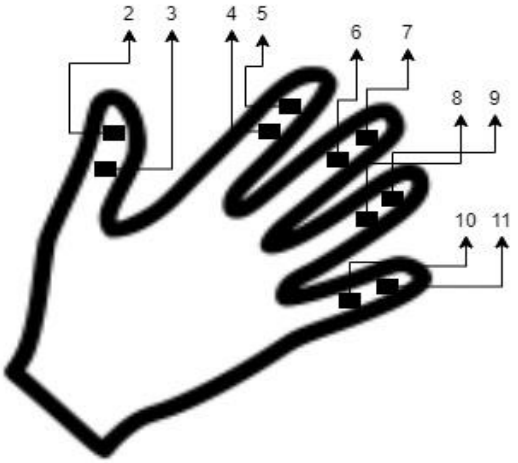

Figure 6: Hand with touch sensors

The touch sensor is made of a 10mm 1 to 11 square copper plate located on the robot arm. This system for recognizing gestures made by the user contains a touch sensor, in which the contact of the electrical property of the user's skin is measured and provides more accurate information, accurate imitation for indicating the gestures of the robot's hand.

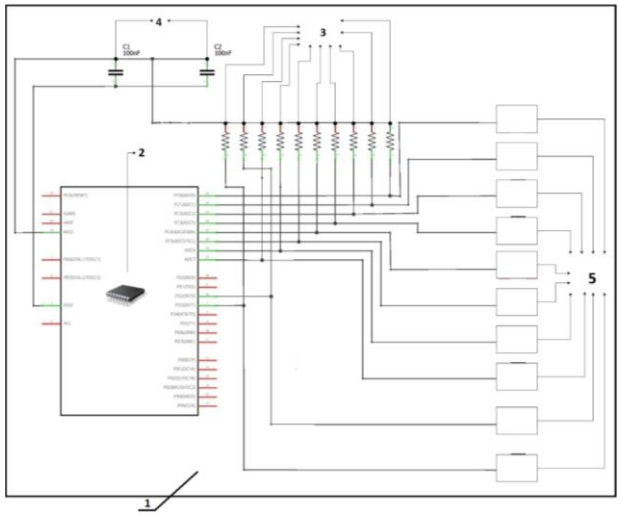

Figure 7: Schematic of a touch sensor

Figure 7 shows the layout of the touch sensor. A microcontroller 1 with touch sensors 2, 3,4,5,6,7,8,9,10,11 is used as a sensor circuit and is connected in parallel to pull-up resistors 12. The resistor is connected to capacitors 13. This sensor detects a touch data.

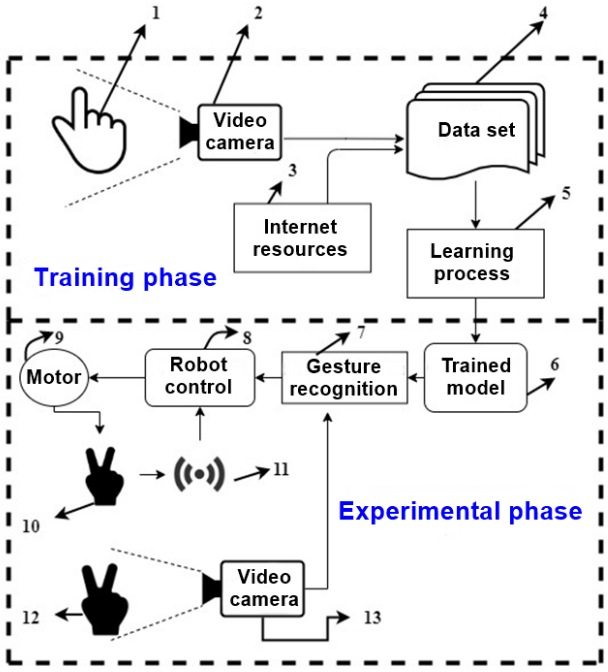

Figure 8. Hand gesture recognition system

Figure 8 shows the hand gesture recognition system. Hand gesture recognition is divided into two phases, the upper part is the learning phase, the lower part is the experiment phase. The hand of a person 1 is shown in the upper part. Next, a hand gesture is taught using a video camera 2, which photographs the position of the hand of the fingers and collects a photo into dataset 4 and simultaneously receives information from Internet resources 3. In dataset 4, 800 were collected for each gesture. images that are kept from different people with different hand sizes. In the process of training model 5 using the gesture dataset by classification, training occurs. After the trained model 6, the model is loaded into gesture recognition 7, in which the process of the experiment phase takes place and the gesture data 12 is received through the camera 13 and fed to the gesture recognition unit 7. The recognized gestures are sent to the robot control unit 8. In the control unit 8, the robot gives a command to rotate the angle of degrees for the servos 9. After that, the robot shows the recognized hand gestures 10. In the recognized hand gestures of the robot 10, touch

sensors 11 are installed. The touch sensor 11 provides the tactile data information to the unit 8. The touch sensor 8 provides more accurate information, an accurate imitation for indicating the gestures of the robot hand.

There are a number of world-class recognition system architectures available to application developers for image classification and image regression. In this study, we used the smallest version of ResNet-18. ResNet is a residual network of building blocks that include "fast connections" that skip one or more layers. The output of the shortcut is appended to the output of the skipped layers. The authors demonstrate that this method facilitates network optimization and provides higher gains in accuracy at significantly increased depths. Featured ResNet architectures range from 18-layer depth to 152-layer depth. For our purposes, the smallest ResNet-18 provides a good balance of performance and efficiency, well suited to the Jetson Nano.

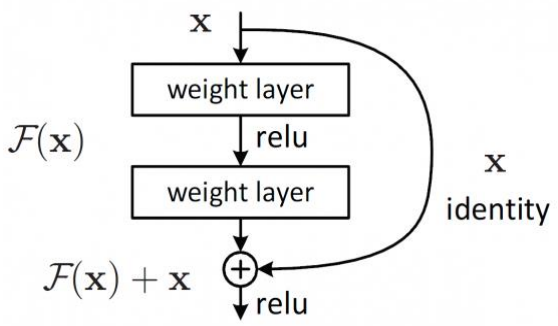

Figure 9: Residual learning: building block

The main element of ResNet is a Residual block (residual block) with a shortcut connection through which data passes unchanged. The Res block is a series of convolutional layers with activations that convert the input signal x to F (x). A shortcut is the identity transformation x → x.

As a result of this design, the Res block teaches how the input signal x differs from F (x). Therefore, if on some layer the network has already sufficiently well approximated the original function generating data, then on further layers the optimizer can make the weights in the Res blocks close to zero, and the signal will pass through the shortcut connection almost unchanged. Figure 9 shows residual learning: the building block. A residual neural network, on the other hand, has short connections parallel to normal convolutional layers. Mathematically, the ResNet layer roughly computes y = f (x) + x. These short paths act like highways and gradients can easily flow back, resulting in faster learning times and a lot more layers. The winner model that Microsoft used in ImageNet 2015 has 152 layers, which is almost 8 times deeper than best CNN.

Calculation method. The range of human speech is 20Hz - 20KHz [3,4,5].

The original speech signal is presented in discrete form as:

$$x[n], 0 \le n < N$$

Applying the Fourier transform to it:

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi i}{N} kn}, \qquad 0 \le k < N$$

Calculation of the frequency value in chalk-scale:

$$B^{-1}(b) = 700(\exp\ (b/1125)\text{-}1)$$

Calculation of signal energy:

$$S[m] = \ln\left(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right), \qquad 0 \le m < M$$

Application of DCT and we obtain a set of signal features, MFCC coefficients:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos\left(\pi n(m + 1/2)/M\right), \qquad 0 \le n < M$$

4 CMUSphinx system

At the moment, the version of sphinx4-5prealpha is actual. Sphinx-4 is a modular framework. The modular structure allows you to vary the parameters of the system based on the requirements of a specific task. There are 3 main modules: FrontEnd, Decoder and Linguist [21].

FrontEnd converts the input to a parameter vector. Linguist builds SearchGraph based on the selected language, acoustic models and vocabulary. Finally, the Decoder's submodule - SearchManager - uses the vector of parameters and the constructed graph for decoding and produces the result [21].

FrontEnd combines several chains of communicating data handlers modules, each of which is capable of producing different parameters

as a result of calculations. The presence of several chains allows both to perform calculations for different types of parameters, and to receive several input signals simultaneously [21].

Linguist. As mentioned above, the construction of SearchGraph in Linguist is based on language data obtained from the language and acoustic model. Each of them represents the HMM for the elementary sound units used in a particular system. The dictionary compares words from a language model and a combination of elements of an acoustic model [21].

The language model describes the structure of language at the word level. The following two kinds of models are commonly used: graph grammars and n-gram models. Graph grammars represent a directed graph in which the vertices are words, and each edge corresponds to a weight, which is the probability of going to the next word. In the case of an n-gram model, there is a set of probabilities to encounter a given word if the previous n − 1 words are known. In this work, we used the trigram language model for American English [27].

The dictionary contains pronunciation options for all words found in a given language model. The pronunciation of a word is divided into sets of some elementary blocks. For example, abandon ↔ AH B AE N D AH N [27].

The acoustic model maps elements of speech (in this case, triphones) to the HMM. Naturally, the display can receive information about the context and position in the word. The left and right contexts were discussed above, and the position information shows whether the triphon is at the beginning, middle, or end of a word. Or it is itself a word [27].

In Linguist, each word is divided into context-dependent elementary blocks (according to information from the dictionary), which are then used to construct a set of HMM graphs using an acoustic model. The vertices of these graphs are phonemes or triphons, and the edges have weights - the probabilities of transition between phonemes). And on the basis of the obtained graphs and the language model, SearchGraph is built [27].

SearchGraph is a directed graph in which each vertex represents an emitting or non-emitting state. The sound features under consideration correspond to the producing states, and the nonproductive states represent higher-level linguistic constructions, such as words or phonemes, which are indirectly related to the calculation of features. The arcs of the graph represent transitions between states, each arc corresponds to a certain probability of passing along it [27].

Decoder. The main function of this module is to obtain a set of hypotheses based on features calculated in FrontEnd and SearchGraph built in Linguist. Decoder sends a request to a submodule called SearchManager to recognize multiple frames with the above data. At each stage of operation, SearchManager builds all paths that reach the final non-productive state. SearchManager uses a token transfer algorithm. For the algorithm being used, the SearchManager may, but does not have to, contain multiple active tokens (ActiveList). To simplify computations, the Pruner submodule abbreviates the set of tokens. The Scorer submodule computes on-demand estimates of the distribution density for the given states at the given times [27].

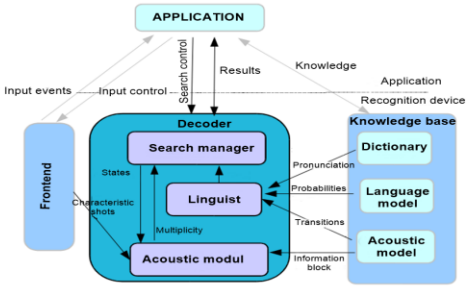

Fig. 10. CMUSphinx architecture

5 Results of gesture recognition

The development of a gesture recognition system was required to start creating a database for the Kazakh sign language - a gesture dictionary. The initial step in this direction was the creation of a database consisting of the dactyl alphabet of forty-two gestures, shown in Figure 11.

Figure 11: Dactyl alphabet of the Kazakh sign language

Dactylology or finger reading requires a certain degree of positioning accuracy. In fig. 12 shows a verbal robot that can represent complete fingerprinting of Kazakh Sign Language. This fingerprint and its results demonstrate how reasonable it is to expect a positive outcome in a robotic hand signature. Since the hand is able to reproduce the complete alphabet, the next step is to test it with deaf and hard of hearing users outside the project to get and appreciate the feedback.

Figure 12: Kazakh fingerprinting is developed by the Inmoov robotic arm to adjust the position of the joint

For each gesture, 8000 samples were collected. These samples were obtained from four different people with different hand sizes. 8000 samples were randomly sorted and then divided into two parts: 5000 samples for simulation and 3000 samples for evaluation. In GMM, component count was a hyperparameter that had to be determined first.

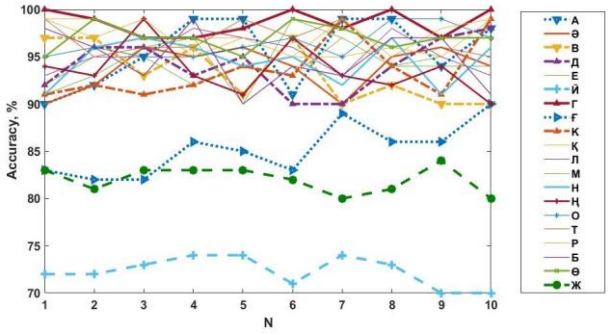

Figure 13: Recognition rate of each gesture (Mi)

As shown in Figure 13, the recognition rate of each gesture (Mi) was plotted. It is clear that K does affect the accuracy of the simulation. Higher K results in higher speed. To balance accuracy and computational complexity, the gesture-based demo platform used a value of K = 3. The average recognition rate was over 98%. The detection time was 3ms on an Intel Core I7-8650U processor @ 1.9 GHz, which is quite enough for this application.

The data for analysis was provided by the Artificial Intelligence and Robotics Laboratory. The dataset consists of 1480 audio recordings from 20 columns with 74–75 recordings. Each audio recording consists of phrases in the Kazakh language with an average duration of 6 seconds. To identify the speaker, the following data were collected: name, gender, place of birth, year of birth [25].

All audio materials have the same characteristics: file extension: wav; digital conversion method: PCM; discrete frequency: 44.1 kHz; bit depth: 16 bit; number of audio channels: one (mono).

Sound and recording of one speaker took an average of 40-50 minutes, including the time required to prepare the speaker, equipment and doublings, which corresponds to 74-75 received files, with a total duration of 7-8 minutes for each speaker [25].

To carry out scientific research and check the performance of the developed algorithms and software, the software simulator of three-dimensional modeling Robo DK was used (Fig. 5-8).

CMUSphinx is implemented in C / C ++ [21]. The intelligent verbal complex was developed in Python with the CMUSphinx communication module and the PyOpenGL graphical command execution simulator. The model of a manipulation

robot for the basics of three-dimensional modeling was chosen by the work of ABB. The graphic results of the study are shown in Fig. 3-6.

The methodology of the work is based on the technique of the average combination of acoustic phonemes for three languages (Kazakh, Russian, English). Fundamentally, the method of work does not differ from each other, as for English, Russian or Kazakh languages [21,22,23]. The aim of the study was to test the methodology and quality of voice recognition in three languages. The obtained WER results showed fluctuations in the range of ~ 61-65%, which is a pretty good result for the first works on language integration.

In the system module for controlling the robot, blocks of programs for recognizing voice commands are implemented in: Kazakh, Russian and English.

The block diagram of the experiment control is shown in Fig. fourteen. The keywords in the control unit database are shown in table 3. The control unit command results are shown in the graph in Fig. 15.

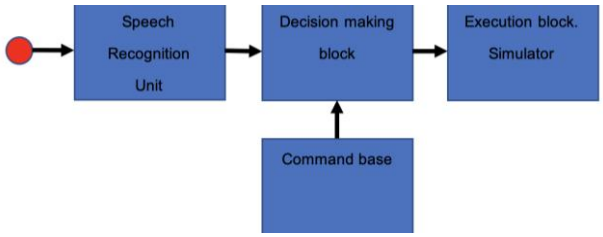

Figure 14: Experiment control block diagram

Table 3. Of basic robot control commands.

№	***Kazakh***	***Russian***	***English***
1	Алға	вперед	go
2	Артқа	назад	back
3	Оңға	направо	right
4	Солға	налево	left
5	Тоқта	стоять	stop
6	Көтеру	поднять	catch up
7	Төмендету	опустить	let go
8	Алу	схватить	catch
9	Қою	положить	put down

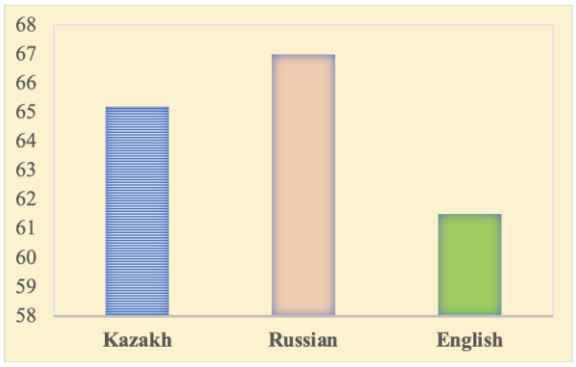

Figure 15: Results of implementing robot control commands

4 Conclusion

This article analyzes the most famous sign languages, the correlation of sign languages, and also considers the development of a verbal robot hand gesture recognition system in relation to the Kazakh language. Within the framework of the system, the speed and accuracy of recognition of each gesture of the verbal robot are calculated. The average recognition accuracy was over 98%. The detection time was 3ms on a 1.9GHz Jetson Nano processor. A complete fingerprint of the Kazakh sign language for a verbal robot is also proposed. To improve the quality of gesture recognition, a machine learning method was used. The operability of the developed technique for recognizing gestures by a verbal robot was tested, and the efficiency of the algorithms and software was evaluated on the basis of computational experiments.

As part of verbal robot arm gesture recognition, natural interaction is achieved through the movement of the robot arm, which is controlled by servo drives. The developed system for recognizing hand gestures by functionality is divided into two phases: the upper part is the learning phase, the lower part is the experiment phase. In the upper part, the process of training the hand gesture recognition model is going on, which is teaching the position of the fingers of the hand, which the robot should learn, in the lower part, the robot loads the trained model and repeats the recognized gestures of the robot hand. For the verbal robot, a complete fingerprint of the Kazakh sign language is presented. This fingerprint and its results demonstrate how reasonable it is to expect a positive result in robotic systems that use hand movements.

Since the hand is able to reproduce the complete alphabet, the next step is to test it with deaf and hard of hearing users to get and appreciate the feedback. A robot with interactive learning capabilities can quickly adapt to different situations as the user can train it specifically for that situation.

Acknowledgement

This work is supported by grant from the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the Work №AP08053034 «Development of new methods for modeling and recognition of Kazakh sign language», Institute Information and Computational Technologies CS MES RK.

References:

[1] Maung, T. H. H. 2009. Real-Time Hand Tracking and Gesture Recognition System Using Neural Networks. Proceedings of World Academy of Science: Engineering & Technology, 50, 466-470.

[2] Mitra, S. & Acharya, T. 2007. Gesture recognition: A Survey. IEEE Transactions on Systems, Man and Cybernetics. IEEE

[3] Bourennane, S. & Fossati, C. 2010. Comparison of shape descriptors for hand posture recognition in video. Signal, Image and Video Processing, 6, 147-157.

[4] Yoon, J.-H., Park, J.-S. & Sung, M. Y. Vision-Based bare-hand gesture interface for interactive augmented reality applications. 5th international conference on Entertainment Computing, September 20-22 2006 Cambridge, UK. 2092520: Springer-Verlag, 386-389

[5] Buchmann, V., Violich, S., Billinghurst, M. & Cockburn, A. 2004. Fingartips: Gesture Based Direct manipulation in Augmented Reality. 2nd international Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia. Singapore: ACM

[6] Trigueiros, P., Ribeiro, F. & Lopes, G. Vision-based hand segmentation techniques for human-robot interaction for real-time applications. In: TAVARES, J. M. & JORGE, R. M. N., eds. III ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, 12-14 de Oubtubro 2011 Olhão, Algarve, Portugal. Taylor and Francis, Publication 31-35.

[7] Vatavu, R.-D., Anthony, L. & Wobbrock, J. O. 2012. Gestures as point clouds: a $P recognizer for user interface prototypes. 14th ACM international conference on Multimodal interaction. Santa Monica, California, USA: ACM.

[8] Wobbrock, J. O., Wilson, A. D. & Li, Y. 2007. Gestures Without Libraries, toolkits or training: a $1 recognizer for user interface prototypes. Proceedings of the 20th annual ACM symposium on User interface software and technology. Newport, Rhode Island, USA: ACM.

[9] Kratz, S. & Rohs, M. 2011. Protractor3D: a closed-form solution to rotation-invariant 3D gestures. 16th International Conference on Intelligent User Interfaces. Palo Alto, CA, USA: ACM.

[10] Kim, T. 2008. In-Depth: Eye To Eye - The History Of EyeToy [Online]. http://www.gamasutra.com. Available: http://www.gamasutra.com/php-bin/news_index.php?story=20975 [Accessed 29-03-2013 2013].

[11] Chowdhury, J. R. 2012. Kinect Sensor for Xbox Gaming. M.Tech CSE, IIT Kharagpur

[12] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H. & Presti, P. 2011. American sign language recognition with the kinect. 13th International Conference on Multimodal Interfaces. Alicante, Spain: ACM.

[13] Ong, S. C. & Ranganath, S. 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans Pattern Anal Mach Intell, 27, 873-91.

[14] Holt, G. A. T., Reinders, M. J. T., Hendriks, E. A., Ridder, H. D. & Doorn, A. J. V. Influence of handshape information on automatic sign language recognition. 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction, February 25-27 2010 Bielefeld, Germany. 2127632: Springer-Verlag, 301-312

[15] Tara, R. Y., Santosa, P. I. & Adji, T. B. 2012. Sign Language Recognition in Robot Teleoperation using Centroid Distance Fourier Descriptors. International Journal of Computer Applications, 48.

[16] Wikipedia. 2012. Língua gestual portuguesa [Online]. Available: http://pt.wikipedia.org/wiki/Lingua_gestual_portuguesa 2013].

[17] Vijay, P. K., Suhas, N. N., Chandrashekhar, C. S. & Dhananjay, D. K. 2012. Recent Developments in Sign Language Recognition : A Review. International Journal on Advanced Computer Engineering and Communication Technology, 1, 21-26.

[18] Chaudhary, A., Raheja, J. L., Das, K. & Raheja, S. 2011. Intelligent Approaches to interact with Machines using Hand Gesture Recognition in

Natural way: A Survey. International Journal of Computer Science & Engineering Survey, 2, 122-133.
[19] Trigueiros, P., Ribeiro, F. & Reis, L. P. A comparison of machine learning algorithms applied to hand gesture recognition. 7th Iberian Conference on Information Systems and Technologies, 20-23 July 2012 Madrid, Spain. 41-46.
[20] Hasanuzzaman, M., Ampornaramveth, V., Zhang, T., Bhuiyan, M. A., Shirai, Y. & H.Ueno. Real-Time Vision-Based Gesture Recognition for Human Robot Interaction. IEEE International Conference on Robotics and Biomimetics, August 22-26 2004 Shenyang, China. IEEE, 413-418.
[21] CMU Sphinx Project by Carnegie Mellon University. http://cmusphinx.sourceforge.net/
[22] Ayazbaev G.M., Kim A.V., Kunelbayev M. Development of software and hardware systems verbal intelligent robot. International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) ISSN(P): 2249–6890; ISSN(E): 2249– 8001 Vol. 10, Issue 3Jun 2020, 1839–1850
[23] Mamyrbayev O., Turdalyuly M., Mekebayev N., Alimhan K., Kydyrbekova A., Turdalykyzy T.. Automatic Recognition of Kazakh Speech Using Deep Neural Networks // Asian Conference on Intelligent Information and Database Systems, 2019.
[24] Kalimoldayev M.N, Akhmetzhanov M., Kunelbayev M., Sundetov T. Information systems of integrated machine learning modules on the example of verbal robot.// News of the national academy of sciences of the RK series of Geology and technical sciences ISSN 2224-5278 Volume 6, Number 438 (2019), 215 – 222p.
[25] Orken Mamyrbayev, Nurbapa Mekebayev, Mussa Turdalyuly, Nurzhamal Oshanova, Tolga Ihsan Medeni, Aigerim Yessentay. Voice Identification Using Classification Algorithms // Intelligent System and Computing, IntechOpen, 2019.
[26] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.
[27] Mikhailov Dmitrii. Analysis of CMUSphinx system performance // Saint-Petersburg 2016 https://robodk.com/