Automation of the translation of Russian sign language into text

Vladimir Bondar
National Research Center «Kurchatov Institute»
Moscow, 123182, Russia
vovabond328@Rambler.ru

***Abstract*—This article is devoted to setting the task of automation of recognition of gestures of Russian language of the deaf. The author designed a prototype of gesture recognition system using a camera with depth sensor.**

***Keywords*—sign language, gesture recognition, depth sensor camera, Kinect**

I. Introduction

In Russia, more than a million people suffer from hearing loss. They are forced to communicate among themselves and with normal hearing people through special sign language. People with no hearing problems usually do not know sign language. This language barrier prevents people with different physical abilities from communicating freely. The only connecting and stable link between hearing and non-hearing people is sign language interpreters. However, their number is not increasing, but only decreasing.

Since most people now have access to a computer, it is logical to use emerging computer technologies to establish communication between hearing impaired people and people with hearing disabilities. [8] In recent years, technical devices and ways of recognizing video images have been improved - there are cameras with depth sensors that can produce three-dimensional images.

Microsoft has developed cameras with the depth sensor Kinect, which allows you to use new features for gesture recognition. Depth sensors have existed for a relatively long time, but Kinect has a number of significant advantages in front of them: large distribution, relatively low cost and availability of RGB camera. The depth map obtained from Kinect is invariant to lighting conditions and background, as it is based on infrared radiation. Only strong fog and some other weather conditions can be an obstacle for this type of recorders.

Thus, there are technical possibilities to develop an automatic translator of Russian sign language into text. As for any natural language processing system, here, in addition to the direct recognition of video information, there are problems of constructing the semantic structure of a sentence, solving problems of polysemy, etc.[1] These questions are beyond the scope of this article. We will focus on gesture recognition and text identification by a recognized gesture.

Microsoft Kinect camera was used as a video input device for gesture information.

II. Concept of Gestural Speech

So, what's a gesture speech? "Gesture speech is a way of interpersonal communication between hearing-impaired people through gestures, characterized by peculiar lexical and grammatical patterns. Gesture speech is used as an additional tool (along with basic verbal speech) in the education and upbringing of children with hearing loss. For official communication, calculating gesture speech is used: the consistent use of gestures to reproduce words." [2] In this case, a gesture (from Latin Gestus - body movement) is understood as "some action or movement of the human body or its part having a certain meaning or meaning, i.e. being a sign or symbol." [3]

The sign language program for people with hearing loss has a complex structure that includes two types of language: spoken sign language, which is used during conversational communication, and calculating sign language, which is used in official communication.

The natural national language is single, the official and ordinary forms differ only in vocabulary, which cannot be said about the language of the deaf, or rather languages, because they are two and use a different set of gestures and different grammar.

Gesture speech is divided into two types: - spoken gesture speech - calculating gesture speech Fingerprinting - alphabetic reproduction of words by using finger configurations. It is used to show the names of own and specific terms. Because it is easy enough to learn, fingerprinting can be a bridge between the deaf and the hearing while the latter are trying to learn spoken sign language. By remembering only a few movements, you can start communicating with the deaf, although on a primitive level.

All gestures in terms of number of movements can be divided into two groups: static and dynamic.

In order to show a static gesture, a simple photo is enough, because it is a fixed position of the body in a certain position. A dynamic gesture (movement) will require animation because it is a sequence of static gestures over a certain period of time.

III. Analysis of gesture recognition tools

A. Microsoft Kinect

Kinect (formerly Project Natal) is a "contactless game controller originally introduced for the Xbox 360 console and much later for Xbox One and Windows PCs. It was developed by Microsoft".

The greatest interest in this work is the 3D depth sensor. The sensor consists of a projector that emits an infrared field and a sensor that reads the field and together they create a depth map. Kinect connects the RGB image and the depth map.

A depth map is an image that stores for each pixel its distance from the camera (instead of color). You can segment the image and find the necessary areas where the gesture is shown using the distance information from the camera.

B. Leap motion

Two monochrome infrared cameras and three infrared radiators are used in Leap Motion to detect the movement of the user's hands. The cameras "scan" the space above the table surface at up to 300 frames per second and transmit the data to a computer where it is processed by proprietary software. Despite their seeming similarity to Microsoft Kinect, these devices are still different.

Leap Motion should be placed under the screen, more precisely - instead of the keyboard (in fact, since it is impossible to do without the traditional control, the controller will have to be placed either behind the keyboard or in front of it as it will be more convenient for the user). As a result, Leap Motion has a rather small scanning area, which, in fact, gets only a small space above the keyboard.

C. 3D camera Intel RealSense

The RealSense camera is an embedded solution for a variety of portable and stationary devices. The camera allows you to get a picture in HD format and provides all the necessary functions of interaction with the person: face and gesture recognition" tracking emotions, highlighting the background, and more.

Simple color camera transmits the picture in the plane. For example, if your left hand is closer to the camera than your right hand, the machine can 'understand' that the hands are at different distances only using special algorithms for pattern recognition. A depth camera solves this problem much more easily, and works like a bat or echo sounder, sending an IR beam and measuring how long it has returned. Knowing the speed of light, it is easy to calculate the distance from the camera to the object.

Thus, gesture recognition tools were considered. The analysis of the tools is shown in Figure 1. For our system was selected device Microsoft Kinect, because it has the necessary characteristics for gesture recognition and has a well-designed software development kit.

Evaluation criterion	Kinect	Leap motion	Intel RealSense	2 cameras
Visibility of key points	Detects up to six people in space and 25 joints of each of them.	Tracks the movement of fingers and hands. Supports 27 degrees of freedom for each hand.	Allows to track up to 76 key points of the face, tracking gestures and position of hands and fingers in the range from 0.2 to 1.2 m. The number of vision points increased up to 22.	none
Range of conditional visibility	With lens support from 0.5 m, without - from 1 m to 5 m.	Up to 1.5 m	Up to 1.2 m	Depends on the cameras' resolution
Sensitivity to light	low	low	low	Depends on the cameras' resolution and quality
Need for additional software	No	No	No	Yes
SDK quality	Good realisation.	Qualitatively realized only for fine motor skills.	It has some defects in operation.	none

Figure 1. Analysis of gesture recognition tools.

IV. Information model for gesture recognition

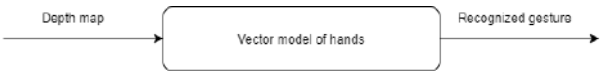

Figure 2. Informational model.

Information from the Kinect camera comes as a depth map. Then, after some processing of the received image, the area of interest is highlighted on it - the human hand [3]. Based on the depth map of the selected part of the image the parameters of the vector model of the hand are calculated.

$$IM =< FV, C, I, O > \quad (1)$$

where IM – Information model;

FV – a vector of hand features. Initial information about the image (color and distance) is analyzed: the position of the hands and the vector of their characteristic features are determined;

C - gesture classifier. According to the found vector of characteristic features of the wrists, it is determined which gesture was shown;

I - set of input values - depth map - distance to each point in the image;

O - set of output values - gesture recognized in the image

V. Gesture recognition system design

The system under design consists of the following modules, which are interconnected with each other:

- image segmentation module;

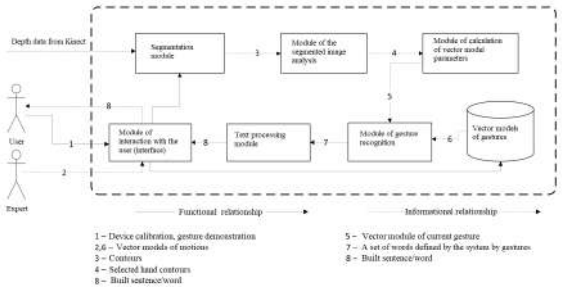

Figure 3. Gesture recognition system design.

- module of the segmented image analysis;
- module of calculation of vector model parameters;
- module of gesture recognition;
- text processing module;
- module of interaction with the user (interface).

VI. General algorithm of gesture recognition system operation

Several works have already been translated from Russian into text, but they focus on extracting information from the hand position. In this work it is suggested to get information about gestures also from facial expressions, as face recognition technologies have stepped very far in the last few years.

The general algorithm of the program consists of the following steps: detect and recognize hands and face on the screen, recognize the gesture and display the text on the screen. The designed system includes two main features: recognize the gesture and display text on the screen.

A. User interface module

The purpose of the module is to provide the user with a convenient interface of working with the program. This includes: displaying the processed image on the screen, the ability to change the settings of the program, displaying the recognized gesture on the screen.

B. Text processing module

The purpose of this module is to form text from recognized gestures, display the formed text on the screen. The recognition of gestures will take place in real time, so the program should recognize the gestures shown, write them to the buffer and display them one after another on the screen. Sentences in the resulting text can be separated by a gesture, which is indicated as a dot.

C. Module of image processing and calculation of vector model parameters

The purpose of image processing modules is to segment the depth map obtained with the Kinect camera, to process the image segmented at the previous stage and to calculate parameters of the vector model of the processed image. Every 1/30 second information from the sensor is supplied to the input, the modules process the input data and calculate the parameters of the vector model of the hand.

Segmentation of the obtained image from the Kinect sensor is necessary for finding the hand or both hands on it. In such cases, as face detection on the image, appearance is a good sign, as eyes, nose and mouth will always be approximately in identical proportions. Therefore, the Haar Cascade method, based on the appearance characteristics of the object, is well suited for facial recognition. In the case of hand recognition, the situation is more complicated: a reliable recognition method can be implemented based mainly on the colour characteristics. Since the colour of the hands can vary depending on the person and context, it seems reasonable to first find the person's face in the image and get information about the colour of the hands based on the colour of the face. The introduced restriction on the presence of a person's face in an image is in any case mandatory, since recognition of a hard language without facial recognition would be unreliable.

Having information about the colour of an object, you should detect it in the image. The task was performed using the Camshift algorithm, the reliability of which has been proved in [Hai et al., 2011]. The model of this algorithm is based on histograms and is trained in the recognition process. Naturally, this algorithm will find all objects of a given color in an image. To prevent it, the information about the distance to the objects in the image is used, i.e. depth map from Kinect sensor.

So, after finding the position (x,y) and dimensions (w,h) of the face in the image using the Haar Cascades method, you can find the average distance to the face using the depth map D:

$$d_f = 1/wh \sum_{i=x}^{x+w} \sum_{i=y}^{y+w} D(i,j) \tag{2}$$

All objects that are closer to the camera than the human face itself can be found using the threshold:

$$D(i,j) > d_f + t_h \tag{3}$$

where t_h– a parameter that determines how close your hands should be to the camera so that the gestures shown can be recognized by the system.

In the second stage of recognition on the segmented image are the contours of human hands with the Canny edge detector.

D. Module of gesture recognition

The purpose of the module is to attribute data obtained from the modules of image processing and calculation of vector model parameters to one of the gestures of the Russian language, embedded in the program. Further it

transfers the recognized gesture to the text processing module. The gesture is recognized, as it was already described earlier, by means of the vector model of a hand. The process is concluded in the following. Originally we receive a vector model from the above described modules. According to this data there are matches, which are stored in the database (matches should not be ideally similar, it is possible to find the most suitable features that will signal the similarity), then formed sets of words obtained from this data.

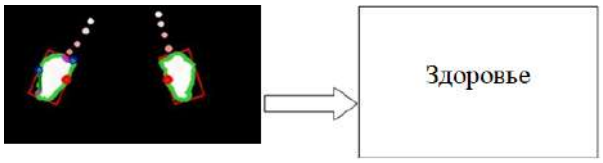

Figure 4. Recognition of "health" gesture.

Thus, the system of translation of Russian sign language into text was designed. This system will be able to translate both static and dynamic gestures.

CONCLUSION

In the course of this study, a gesture recognition system for Russian gestures was designed to recognize gestures by obtaining information about the position of the hands and facial expressions.

As noted, the sign language consists of a number of channels for the transmission of information, so face recognition and hand gestures do not solve the problem of recognition of Russian sign language completely, but is an important element of the future full system. The subject of further research is the implementation of this project as a software product.

In the process of image analysis and text acquisition in natural language semantic technologies of intellectual systems will be used.

REFERENCES

[1] Popov M.YU., Fomenkov S.A., Zaboleeva-Zotova A.V. Principy raboty i osobennosti realizacii universal'nogo semantiko-sintaksicheskogo analizatora russkogo yazyka[Principles of work and features of implementation of the universal semantic-syntactic analyzer of the Russian language]. Vestnik komp'iuternykh i informacionnykh tekhnologii [Herald of computer and information technologies], 2005, No 7 (13), pp. 37-42

[2] Zaboleeva-Zotova A.V., Orlova YU.A., Rozaliev V.L., Fedorov O.S. Predstavlenie harakternyh emocional'nyh zhestov i dvizhenij cheloveka v vide nechyotkih posledovatel'nyh temporal'nyh vyskazyvanij [Presentation of characteristic emotional gestures and movements of a person in the form of vague consecutive temporal expressions],Obozrenie prikladnoj i promyshlennoj matematiki [Overview of Applied and Industrial Mathematics]. 2011, vol 18, no 4, pp. 537-544

[3] Dorofeev N.S., Rozaliev V.L., Zaboleeva-Zotova A.V. Sistema raspoznavaniya zhestov russkogo yazyka gluhih [The system of recognition of gestures of Russian language of deaf people], Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2013, pp. 399-402.

[4] Petrovskij A.B., Zaboleeva-Zotova A.V., Bobkov A.S. Formalizovannoe opisanie dvizhenij cheloveka na osnove nechetkih temporal'nyh vyskazyvanij [Formalized description of human movements based on vague temporal expressions] Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte VII-ya Mezhdunarodnaya nauchno-prakticheskaya konferenciya, k 80-letiyu Dmitriya Aleksandrovicha Pospelova: sbornik nauchnyh trudov v 3 tomah. [Integrated models and soft computations in artificial intelligence]. 2013. pp. 488-495.

[5] Rozaliev V.L., Zaboleeva-Zotova A.V. Methods and models for identifying human emotions by recognition gestures and motion Journal of Politics. 2013, vol. 2. p. 67.

[6] Konstantinov V.M., Rozaliev V.L., Orlova Y.A., Zaboleeva-Zotova A.V. Development of 3D human body model Advances in Intelligent Systems and Computing. 2016. vol. 451. pp. 143-152.

[7] Vybornyi A.I., Rozaliev V.L., Orlova Yu.A., Zaboleeva-Zotova A.V., Petrovsky A.B. Controlling the correctness of physical exercises using microsoft kinect Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh system [Open semantic technologies for intelligent systems], 2017, Iss. 1, pp. 403-406.

[8] Orlova Yu.A. Modeli i metody informacionnoj podderzhki kommunikacii lyudej s ogranichennymi vozmozhnostyami [Models and methods of information support for communication of people with disabilities.]. – 2016

[9] Sean Kean, Jonathan Hall, Phoenix Perry, Meet the Kinect: An Introduction to Programming Natural User Interfaces (Technology in Action). Apress, 2011. – 220 c.

[10] H. Hai, L. Bin, H. BenXiong and C. Yi, "Interaction System of Treadmill Games based on depth maps and CAM-Shift", IEEE 3rd International Communication Software and Networks (2011), pp.219-222

Автоматизация перевода русского жестового языка в текст

Бондарь В.В.

Данная статья посвящена постановке задачи автоматизации распознавания жестов русского языка глухих. Автором спроектирован прототип системы распознавания жестов с использованием камеры с сенсором глубины.