Control System of Collaborative Robotic Based on the Methods of Contactless Recognition of Human Actions

Aleksandr Zelensky[1], *Marina* Zhdanova[1,2,3*], *Viacheslav* Voronin[1,2], *Andrey* Alepko[1,3], *Nikolay* Gapon[2], *K.O.* Egiazarian[4] and *Oksana* Balabaeva[2]

[1]Moscow State Technological University "STANKIN", RU-127055, Moscow, Russia
[2]Don Sate Technical University, RU-344000, Rostov-on-Don, Russia
[3]Scientific-manufacturing complex “Technological centre”, RU-124498, Zelenograd, Russia
[4]Tampere University of Applied Sciences, FI-33720, Tampere, Finland

Abstract. Human-robot collaboration is a key concept in modern intelligent manufacturing. Traditional human-robot interfaces are quite difficult to control and require additional operator training. The development of an intuitive and native user interface is important for the unobstructed interaction of human and robot in production. The control system of collaborative robotics described in the work is focused on increasing productivity, ensuring safety and ergonomics, minimize the cognitive workload of the operator in the process of human-robot interaction using contactless recognition of human actions. The system uses elements of technical vision to get of input data from the user in the form of gesture commands. As a set of commands for control collaborative robotic complexes and training the method proposed in the work, we use the actions from the UTD-MHAD database. The gesture recognition method is based on deep learning technology. An artificial neural network extracts the skeleton joints of the human and describes their position relative to each other and the center of gravity of the whole skeleton. The received descriptors feed to the input of the classifier, where the assignment to a specific class occur. This approach allows reducing the error from the redundancy of the data feed at the input of the neural network.

1 Introduction

In recent years, human-robot collaboration (HRC) has become a key technology of intelligent manufacturing. Conception HRC allows combining human-operator and the robot in one workspace over common tasks, instead of separating the responsibilities of them for safety reasons. These robots are called collaborative robots or cobots. According to the international standard ISO/TS 15066:2016 [1], cobots are robots designed to work together with people within a specific workspace, equipped with visual and speech recognition.

HRC combines the interaction of human and robot into a single intelligent system, and as a result, effectively organize flexible methods of automation and reconfiguration of production processes [2]. Human interaction with robotic systems reaches a new level and strives to become similar to human-human communication due to the development of speech, image, and video processing technologies.

Currently, human-robot interaction interfaces can be divided into two categories:

- remote interfaces using gestures and voice;
- physical interfaces, such as pendant, teach pendant, tactile interfaces, etc.

Teach pendant is the primary interface between humans and robots in traditional automated manufacturing. This control device is an information screen and a specific set of buttons for control. Generally, the pendant has an intricate layout of buttons to provide a wide range of functionality, and the graphical interface varies among different manufacturing companies. It is concerning, so any operation of the robot requires additional operator training.

Gesture control may assume the use of additional contact controls, such as gloves, bracelets, and other body sensors. These devices track the position of the user's hand in real-time and provide the ability to send a control command to the robot through natural movements.

However, gloves prevent the user from moving and are uncomfortable to wear. Also, using that device is often limited in an industrial context, because gloves and motion capture devices often require calibration before use, which significantly increases preparation time. Wearable devices also require frequent maintenance due to wear, which increases the cost.

The rapid development of technical vision technologies and their implementation in various manufacturing sectors makes it relevant to use such technologies in the development of human-robot interaction systems using non-contact recognition of human actions.

[*] e-mail: mpismenskova@mail.ru

Over the past decade, many studies have been presented on the methods of non-contact recognition of human actions. The variety and complexity of recognition problems, the specificity of the features, and the actions do not allow one universal approach to solving them to be implemented. Most methods have several drawbacks and require significant time and computational resources.

Many methods consider only the contents of the frames and do not take into account the time features of the video.

For practical purposes, methods implemented based on neural networks are especially popular [3-6]. In some situations, recognition is performed by minimizing the distance to the standard (the principle of nearest neighbors). Algorithms that perform the analysis of objects' features have acceptable results for simple tasks, but the recognition of three-dimensional scenes requires exceptionally large computing power, and the quality of the results is insufficient.

Spatio-temporal “points of interest” or “local features” are those points where the local neighborhood has significant differences, both in the spatial and temporal domains. Most local space-time descriptors are extensions of functions built based on singular points of two-dimensional space into a three-dimensional region. These methods record changes in motion in spatial and temporal dimensions in the surround of points of interest [7]. Methods based on calculations of the spatio-temporal local features have the stability to scaling, to the weak rotation of the image. There are difficulties when extracting the Spatio-temporal dependence, instability to images having a complex structure, and overflowing background.

A global descriptor is a feature vector obtained by analyzing the entire image as a whole. Typically, in such methods, each point in the image contributes to the descriptor value. One of the methods for constructing descriptors for a video sequence is the global video descriptor (Global Video Descriptor) [8]. This approach works well with video sequences containing simple sets of actions. The method's performance decreases when occlusions occur, the background is full, or several objects appear in the frame.

Texture features such as LBP-TOP [9-11] are used in the recognition of human actions. The use of that descriptors requires an additional pre-processing step, which increases the computational cost.

The control system based on technical vision is non-contact in nature, which means that it is not invasive for the movement of the user and the process associated with it. Also, the current state of such systems can support tracking the position of the human body, which allows you to execute gesture commands or enter body movements. To make full use of human skills when interacting with a collaborative robot, it is necessary to provide informative input data for the robot and intuitive commands for the operator. Thus, some collaborative robotic system should be convenient for the operator and allow the novice user to interact without any expert knowledge.

The development of a new system of human-robot interaction using gestures or speech, or augmented reality, is a very relevant task and allows to avoid the disadvantages of traditional means of interaction, such as keyboards, mice, touch pendant and learning panels.

The purpose of this work is to develop a control system for collaborative robotic complexes to increase productivity, safety, and ergonomics in the process of human-robot interaction using non-contact recognition of human actions.

2 Robot control system

The developed control system for the robotic complex is an interface of interaction between a person and a robot, which receives commands from the operator in the form of gestures. The system uses elements of technical vision to contactless receive input from the user.

The architecture of the robot control system is shown in Figure 1, which consists of an industrial robot, a data processing unit, an IP stack (TCP/UDP), and a robot control controller.

The data processing unit consists of software (software). It is a central user interface control element that receives input signals from visual information sensors and generates input data for a decoder program in the robot controller. The controller receives a signal from the software of the data processing unit via the IP stack and starts an industrial robot to perform actions.

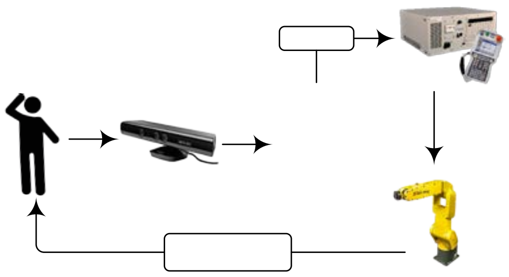

Fig. 1. Robot control system.

It is supposed to use a specialized microcontroller with the RISC-V computing core optimized for vector and floating-point operations as a hardware platform. This microcontroller is under development at the State Engineering Center “MST” STANKIN” jointly with the Scientific-manufacturing complex “Technological centre”.

When we work with neural networks, especially deep ones, our network itself can occupy hundreds of megabytes. For example, the memory requirements of object detection networks are as follows:

model	input size	param memory	feature memory
rfcn-res50-pascal	600 x 850	122 MB	1 GB
rfcn-res101-pascal	600 x 850	194 MB	2 GB

ssd-pascal-vggvd-300	300 x 300	100 MB	116 MB
ssd-pascal-vggvd-512	512 x 512	104 MB	337 MB
ssd-pascal-mobilenet-ft	300 x 300	22 MB	37 MB
faster-rcnn-vggvd-pascal	600 x 850	523 MB	600 MB

In the case of using ASIC, which in essence is an accelerator combined with the RISK-V core, there are the following advantages and disadvantages:

Pros:

1. The lowest chip cost compared to all previous solutions.
2. Lowest power consumption per unit of operation.
3. Quite high speed (including, if desired, a record).

Minuses:

4. Very limited options for updating the network and logic.
5. Highest development cost compared to all previous solutions.
6. Using ASIC is cost-effective mainly for large runs.

The choice of this architecture is due to the presence of an open set of instructions, which already contains, as an extension, tools for working with vector data, such as 32 separate vector registers v0 - v31, which are scalable sections of memory (Figure 2).

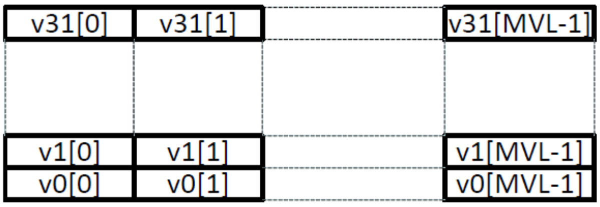

Fig. 2. The structure of vector registers.

In addition, instructions for working with vector registers, such as vmul, vadd, are added in this extension.

The open architecture allows the use of such advanced instruction sets and add specialized hardware computers (Figure 3).

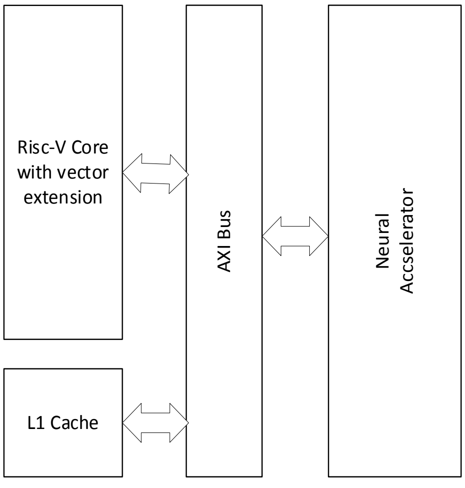

Fig. 3. Structure of the interaction of a neural accelerator and the RISC-V core.

Microcontrollers of the RISC-V family are characterized by scalable architecture, which in some cases allows them to approach the DSP microprocessors in computational capabilities.

3 The human gesture recognition algorithm

To generate input commands to the robot controller, an algorithm for recognizing the actions of human gestures based on previously prepared high-level skeleton data based on a neural network is developed. The use of human skeleton data can reduce the error of redundant information.

The coordinates of the three-dimensional joints of the human-operator are calculated based on the processing of depth sensor data. These include the neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, and left wrist (Figure 4).

To prepare the data at the input of the neural network, coordinates are normalized with respect to body length and relative to the center of gravity. The method of recognizing actions based on the construction of the human skeleton is described in detail in [13–15].

The length of the body (6) is calculated as the sum of the lengths of the individual parts: head (1), torso (2), leg from the thigh to the knee and from the knee to the ankle (3, 4), the maximum length of the right or left leg (5) is selected:

$$Length_{head} = \sqrt{(j_{10}(x) - j_9(x))^2 + (j_{10}(y) - j_9(y))^2} \quad (1)$$

$$Length_{torso} = \sqrt{(j_8(x) - j_7(x))^2 + (j_8(y) - j_7(y))^2} \quad (2)$$

$$\mathrm{Length}_{\mathrm{leg\_right}} = \sqrt{(\mathrm{j}_3(\mathrm{x}) - \mathrm{j}_2(\mathrm{x}))^2 + (\mathrm{j}_3(\mathrm{y}) - \mathrm{j}_2(\mathrm{y}))^2} + \sqrt{(\mathrm{j}_2(\mathrm{x}) - \mathrm{j}_1(\mathrm{x}))^2 + (\mathrm{j}_2(\mathrm{y}) - \mathrm{j}_1(\mathrm{y}))^2} \quad (3)$$

$$\mathrm{Length}_{\mathrm{leg}_{\mathrm{left}}} = \sqrt{(\mathrm{j}_4(\mathrm{x}) - \mathrm{j}_5(\mathrm{x}))^2 + (\mathrm{j}_4(\mathrm{y}) - \mathrm{j}_5(\mathrm{y}))^2} + \sqrt{(\mathrm{j}_5(\mathrm{x}) - \mathrm{j}_6(\mathrm{x}))^2 + (\mathrm{j}_5(\mathrm{y}) - \mathrm{j}_6(\mathrm{y}))^2} \quad (4)$$

$$\mathrm{Length}_{\mathrm{leg}} = \max(\mathrm{Length}_{\mathrm{leg\_right}}, \mathrm{Length}_{\mathrm{leg\_left}}) \quad (5)$$

$$\mathrm{Length}_{\mathrm{body}} = \mathrm{Length}_{\mathrm{head}} + \mathrm{Length}_{\mathrm{torso}} + \mathrm{Length}_{\mathrm{leg}} \quad (6)$$

where $j_n(x)$ –the x-coordinate of the n-joint, $j_n(y)$ – the y-coordinate of the n-joint, $Length_{head}$ is the length of the head, $Length_{torso}$ is the length of the torso, $Length_{leg\_right}$ is the length right leg, $Length_{leg\_left}$ - the length of the left leg, $Length_{leg}$ - the maximum length of the leg, $Length_{body}$ - the length of the whole body.

The center of gravity is calculated by the formula (7, 8):

$$Centr_x = \sum_n \frac{j_n(x)}{16} \quad (7)$$

$$Centr_y = \sum_n \frac{j_n(y)}{16} \quad (8)$$

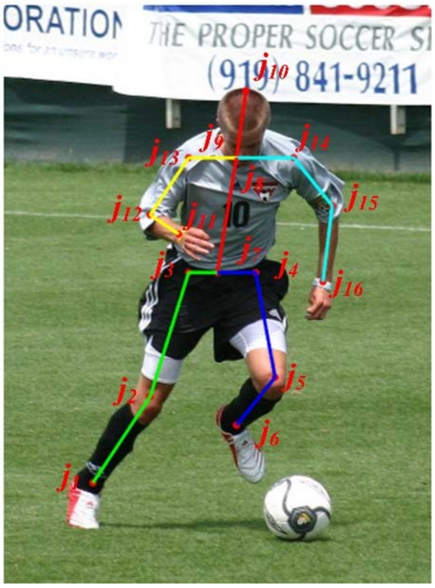

Fig. 4. An example of building a skeleton for an image from the Leeds Sports Pose Dataset test kit [12].

At the next stage, the coordinates are normalized with respect to the body length and the center of gravity (9, 10).

$$i_n(x) = \frac{j_n(x) - Centr_x}{Length_{body}} \quad (9)$$

$$i_n(y) = \frac{j_n(y) - Centr_y}{Length_{body}} \quad 10)$$

where $i_n(x)$ – the x–normalized coordinate of the n–joint, $i_n(y)$ – the y–normalized coordinate of the n – joint.

Also, information about the distance between certain joints, which can characterize the features of the presented posture, namely, the distance from the wrist to the shoulder (11,12) and from the ankle to the hip (13,14), is also fed to the neural network entrance.

$$Dist_{shoulder-wrist-r} = \sqrt{(j_{13}(x) - j_{11}(x))^2 + (j_{13}(y) - j_{11}(y))^2} \quad (11)$$

$$Dist_{shoulder-wrist-l} = \sqrt{(j_{14}(x) - j_{16}(x))^2 + (j_{14}(y) - j_{16}(y))^2} \quad (12)$$

$$Dist_{hip-ankle-r} = \sqrt{(j_3(x) - j_1(x))^2 + (j_3(y) - j_1(y))^2} \quad (13)$$

$$Dist_{hip-ankle-l} = \sqrt{(j_4(x) - j_6(x))^2 + (j_4(y) - j_6(y))^2} \quad (14)$$

The data described above is fed to the input of a neural network, which has 34 inputs. Since 8 types of actions were used to test the network, the output layer is represented by 8 neurons.

During the experiments, various configuration options for the neural network were considered. The best result was achieved by using a perceptron consisting of 34 neurons on the input layer (logsig activation function), 62, 32 and 16 neurons on three hidden layers (logsig activation function) and three neurons on the output layer (linear activation function).

4 Investigation of the human gesture recognition algorithm

The developed human gesture recognition algorithm is tested on the UTD-MHAD dataset [16]. The presented modification of the neural network allows you to get the correct recognition result in 85.5% of cases.

5 Conclusion

It was developed a control system for collaborative robotic systems to increase productivity, safety, and ergonomics in the process of human-robot interaction based on contactless recognition of human actions. In developing the gesture recognition method, deep learning technology was used to identify the main points of the skeleton, and to analyze their relative positions, which allows tracking multiple hypotheses for various gesture recognition scenarios in the interaction of a person and a robot.

Acknowledgments. The study has been supported of the Russian Foundation for Basic Research (RFBR) (projects No. 18-313-00090 and No. 18-37-00101). This work was carried out using equipment provided by the Center of Collective Use of MSUT "STANKIN".

References

1. ISO/TS 15066:2016, “Robots and robotic devices - collaborative robots,” International Organization for Standardization, Standard ISO/TS 15066:2016 (2016)
2. Y. Koren, U. Heisel, F. Jovane, T. Moriwaki, G. Pritschow, G. Ulsoy, & H. Van Brussel, CIRP annals, **48**(2), 527-540 (1999)
3. F. Caba Heilbron, V. Escorcia, B. Ghanem, & J. Carlos Niebles, CVPR, 961-970 (2015)
4. L. Wang, Y. Xiong, D. Lin, & L. Van Gool, CVPR, 4325-4334 (2017)
5. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, & T. Darrell, CVPR, 2625-2634 (2015)
6. C. Liu, J. Liu, Z. He, Y. Zhai, Q. Hu, & Y. Huang, Pattern recognition, **59**, 213-224 (2016)
7. H. Liu, Z. Ju, X. Ji, C.S. Chan, and M. Khoury, Human Motion Sensing and Recognition, 233-250 (2017)
8. B. Solmaz, S.M. Assari, M. Shah, Machine vision and applications, **24** (7), 1473-1485 (2013)
9. G. Zhao, M. Pietikainen, IEEE transactions on pattern analysis and machine intelligence, **29**(6), 915-928 (2007)
10. V. Kellokumpu, G. Zhao, M. Pietikäinen, BMVC **1** (2008)

11. R. Mattivi, L. Shao, International Conference on Computer Analysis of Images and Patterns, 740-747 (2009)
12. S. Johnson, M. Everingham. BMVC, **2**(4) (2010)
13. M. Pismenskova, O. Balabaeva, V. Voronin & V. Fedosov, MATEC Web of Conferences (EDP Sciences, Rostov-on-Don, 2017) **132**, 05016
14. A.A. Zelensky, V.A. Franz, Vestnik MSTU Stankin, **3**(46), (2018) [in Russian]
15. A.A. Zelensky, M.M. Pismenskova, Vestnik MSTU Stankin, **3**(46), (2018) [in Russian]
16. C. Chen, R. Jafari, N. Kehtarnavaz, IEEE International Conference on. IEEE, 168-172 (2015)
17. S.N. Grigoriev, G.M. Martinov, Proc. CIRP, **41**, 858-863 (2016)
18. S.N. Grigoriev, M.P. Kozochkin, F.S. Sabirov, and A.A. Kutin, Proc. CIRP, **1**, 599-604 (2012)
19. S.N. Grigoriev, V.A. Sinopalnikov, M.V. Tereshin, and V.D. Gurin, Measur. Techn., **55**(5), 555-558 (2012)
20. S.N. Grigoriev, V.D. Gurin, M.A. Volosova, and N. Y. Cherkasova, Materialwiss. Werkstofftech., **44**(9), 790-796 (2013)
21. S.N. Grigoriev, G.M. Martinov, Proc. CIRP, **14**, 517-522 (2014)
22. A.G. Ivakhnenko, V.V. Kuts, O.Y. Erenkov, E.O. Ivakhnenko, A.V. Oleinik, Russ. Eng. Res., **37**(10), 901-905 (2017)