Middle-East Journal of Scientific Research 20 (12): 2177-2183, 2014
ISSN 1990-9233
© IDOSI Publications, 2014
DOI: 10.5829/idosi.mejsr.2014.20.12.21105

Distributed System of Real Time Head Gesture Recognition in Development of Contactless Interfaces

[1]Andrey Ostroukh, [2]Viacheslav Nikonov, [2]Irina Ivanova,
[2]Tatiana Morozova and [3]Valeriy Strakhov,

[1]Department "Automated Control Systems",
Moscow Automobile and Road construction State Technical University (MADI),
Moscow, Russia
[2]Department "Automated Control Systems",
Moscow State University Instrument Engineering and Computer Science (MGUPI),
Moscow, Russia
[3]LLC "NEOKOM," Sergiev Posad, Russia

Abstract: The article discusses the gesture recognition system based on head distributed multi-agent network. A model of head movement in 3D space. Head position described by the matrix configurations. Developed a method of recognition of fuzzy gestures, based on recognition of the graph constructed on the basis of fuzzy nodes. Graph is generated for each dynamic gesture, perfect head. It is shown that the recognition process can be performed by a fuzzy neural network. Software agents implement recognition system.

- Presented work distributed recognition system in real time.
- A general scheme for the formation of classifying gestures. The estimation of the volume of data in the form of a matrix.
- Presented data rate in a distributed system.
- The data capacities of the most common data channels.
- Choose the method of transmission of a video stream.

Key words: Fuzzy pattern recognition method gestures · 3D model · A distributed system · Computer vision

INTRODUCTION

Information technologies introduction in industrial enterprises steadily alters the requirements for a "human operator" in automated "man – machine" mechatronic and robotic systems. Volumes of processed information, its polyalternativeness, abundance of diversified sources, means and channels of its receipt and transmission offer this person the challenge of not only instant decision-making based on data on changes in the situation, but also performing a significant number of mainly manipulative operations with electronic equipment in a very short period of time.

The current level of information technology development provides for creation of the so-called contactless interfaces based on the integration of "computer vision" and "voice control" technology. However, such problem as "computer vision" is interconnected with the lack of intellectual system [1-3] integrating diverse information about the environment

Corresponding Author: Andrey Ostroukh, Department "Automated Control Systems",
Moscow Automobile and Road construction State Technical University (MADI), Moscow, Russia.

and its changes. Operators of modern computer systems are still constrained in terms of their “convenient” management because of multiplicity and fragmentation of the channels through which the management is carried out.

In some cases, especially in hazardous industries or extreme emergency, the universal contactless interfaces are essential that do not require the use of input devices for hands or any contact elements. “Taking the load off the operator's hands” – is one of the most important technological challenges of our time.

Head Movement Model in 3d Space: In the imaging process the vertices coordinates are subject to certain transformations. The defined normal vectors undergo such transformations.

Initially, the camera is at the origin of coordinates and is directed along the negative direction of the *OZ* axis. The head is recognized in the video stream by the method of active external appearance models. As a result of the recognition, a 3D model of the head is obtained. The head is regarded as a rigid body with six degrees of freedom.

The head position can be set using the configuration matrix. The configuration matrix is as follows:

$$C=\left(\Theta,\varphi,\psi,t_x,t_y,t_z\right)$$

where *Θ*,□, *ψ* – Euler angles, t_x,t_y,t_z – parallel translation parameters.

Parallel translation – is moving of each point of the object model for some distance in a given direction. Parallel translation matrix T is given by the following formula:

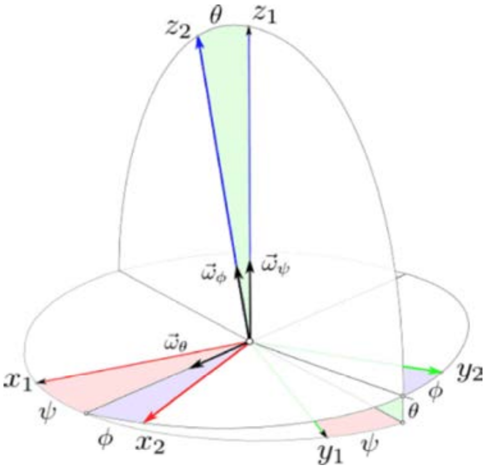

Fig. 1: Movement by the Euler angles

$$\mathrm{T}=\begin{pmatrix}1 & 0 & 0 & t_x\\ 0 & 1 & 0 & t_y\\ 0 & 0 & 1 & t_z\\ 0 & 0 & 0 & 1\end{pmatrix} \qquad (1)$$

where t_x,t_y,t_z – define the displacement vector

Rotation – circle movement of each point on the object model around a certain axis of rotation. The rotation matrix R 4×4 in the general case is as follows:

$$\mathrm{R}=\begin{pmatrix}r_{11} & r_{12} & r_{13} & 0\\ r_{21} & r_{22} & r_{23} & 0\\ r_{31} & r_{32} & r_{33} & 0\\ 0 & 0 & 0 & 1\end{pmatrix} \qquad (2)$$

Rotation can be parameterized in several ways. One of the most common is the use of Euler angles, i.e. rotation decomposition by rotation around three different axes. For example, the R matrix can be expressed as the multiplication of three matrices R_y, R_z, R_x:

$$R(\Theta,\varphi,\psi)=R_y(\Theta)R_z(\varphi)R_x(\psi)$$

where matrices R_y, R_z, R_x respectively define a successive rotation around Y_e, X_e, Z_e axes and are defined by the formulas:

$$R_y(\theta)=\begin{pmatrix}cos\theta & 0 & \sin\theta & 0\\ 0 & 1 & 0 & 0\\ -\sin\theta & 0 & \cos\theta & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$$

$$R_z(\varphi)=\begin{pmatrix}cos\varphi & -\sin\theta & 0 & 0\\ \sin\varphi & \cos\theta & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$$

$$R_x(\psi)=\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & \cos\psi & -\sin\psi & 0\\ 0 & \sin\psi & \cos\psi & 0\\ 0 & 0 & 0 & 1\end{pmatrix}$$

In more detail, the multiplication of these matrices can be written as follows:

$$R(\theta,\varphi,\psi)=\begin{pmatrix} \cos\theta\cos\theta & \begin{matrix}-\cos\theta\sin\varphi\cos\psi\\ +\sin\theta\sin\psi\end{matrix} & \begin{matrix}\cos\theta\sin\varphi\sin\psi\\ +\sin\theta\cos\psi\end{matrix} \\ \sin\theta & \cos\varphi\cos\psi & -\cos\varphi\sin\psi \\ -\sin\theta\cos\varphi & \begin{matrix}\sin\theta\sin\varphi\\ +\cos\theta\sin\psi\end{matrix} & \begin{matrix}-\sin\theta\sin\varphi\sin\psi\\ +\cos\theta\cos\psi\end{matrix} \end{pmatrix}$$

In practice, it is convenient to combine the translation (1) and rotation (2) operations in a single M matrix (4×4)

$$M = TR = \begin{pmatrix} R' & T' \\ 0_{1\times 3} & 1 \end{pmatrix} \qquad (3)$$

where, $T'=(t_x,t_y,t_z)$ – parallel translation vector from formula (1), which defines the displacement from the reference point; $R'=\begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}$ – 3×3 order submatrix of formula (2), $0_{1\times 3}$- row vector consisting of zeros.

Fuzzy Gestures Recognition Method: Fuzzy gestures recognition method is based on recognition of the graph constructed on the basis of fuzzy nodes. The graph is generated for each dynamic head gesture. For collection of statistical distribution of the gesture images it is repeated several times and the path of each repetition is fixed. The number of repetitions is 10-20. For example, the paths of L-gesture are shown in Fig. 2. In this case, to generate the gesture path the computer vision algorithms are used to capture and track the head movement in a video stream: finding a moving object in the frame (frame subtraction algorithm [4]).

L-gesture example in space can be described by the M matrix (3).

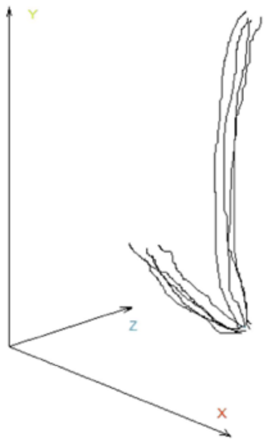

Fig. 2: L-gesture paths

The M matrix produces the unambiguous head position in space at the current time. For example, assume that the user sits in front of the camera at a distance of 0.8 m. Then, the M_0matrix at the initial time will be as follows:

$$M_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0{,}8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The next time when head moving down 20°, M_1 matrix will be as follows:

$$M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0{,}94 & -0{,}34 & 0 \\ 0 & 0{,}34 & 0{,}93 & 0{,}8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Further, if a person turns their head 10° to the right, M_2 matrix will be as follows:

$$M_1 = \begin{pmatrix} 0{,}98 & 0 & -0{,}17 & 0 \\ -0{,}06 & 0{,}94 & 0{,}34 & 0 \\ 0{,}16 & -0{,}34 & 0{,}93 & 0{,}8 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The generalized movement path for L-gesture of all movements can be represented as a graph shown in Figure 3. A_1 vertex of the graph combines a set of points belonging to the path origins, A_2 vertices correspond to the paths inflexion point, A_3 vertex combines all paths and the arcs of the graph indicate the object movement direction along the paths. This graph can serve as a basis for fuzzy gesture model constructing. Each vertex of the graph combines the characteristic points with a certain resemblance. The set of points belonging to the same vertex constitute a cluster. Each point generally belongs to the *m*-dimensional space and represents a set of values of y_1, y_2,..., y_m characteristic features. To determine the clusters fuzzy neural networks are used. At the beginning, it is necessary to describe class instances in a fuzzy representation. Then, for each *c* class, *c* = 1, 2,...*C* a fuzzy model of *c* class is constructed. When a *u* unknown object (head position) must be recognized, the *u* fuzzy representation is compared with each *c* fuzzy model by defining a similarity degree. An unknown object is considered to be recognized in the event that it belongs to the class with the highest similarity degree.

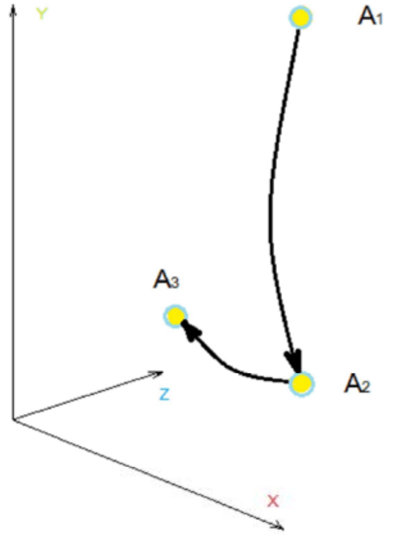

Fig. 3: L-gesture graph

The recognition process can be carried out using fuzzy neural network.

When using ANFIS (AdaptiveNeuro–Fuzzy Inference System), this type of network can be guided by self-organization algorithms. Such algorithms as WTM ("WinnerTakesMost") are used for guiding, where – apart from the winner – the neurons of its immediate surrounding clarify their weights. At the same time, the more distant a neuron is from the winner, the less change occurs to its weights. The weights vector alignment can be defined by generalized dependence, which is represented as follows

$$w_i \leftarrow w_i + \alpha G(i,x)[x - w_i]$$

for all neurons located in the vicinity of the winner, where

w_i – weight factor of the i-th neuron.
$G(i,x)$ function is defined as follows

$$G(i,x) = \{1\, for\ i = I, 0\, for\ i \neq I\}$$

Where I denotes the number of the winner, then we obtain the classical *WTA* algorithm. There are many versions of the *WTM* algorithm, differing primarily in the form of G (i, x) function.

Distributed Recognition System Operation in Real Time: Concurrency of neurons operation in the network can be translated in a distributed system based on software agents [9 – 15]. The term "agent" (Fig. 4) originates in the field of artificial intelligence and describes a logical entity that has certain autonomy in its environment or in its host. The agent can move between hosts. In the computer context, the agent means the entity that combines data, code and is capable of moving between different runtime environments. Agents can provide several advantages, such as reduction of traffic transmitted over a network, decentralization, high reliability and fault tolerance, as well as ease of deployment.

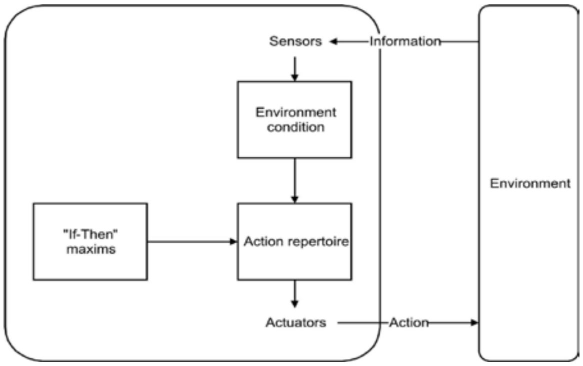

Fig. 4: Software agent structure

The general scheme of the gestures classifying system formation is as follows. There is a population of agents, each of which is characterized by its neural network structure described above. Within each generation, each agent goes through the stage of teaching and testing for each of the built "teaching selection – test selection" pairs. The teaching consists in the fact that the agent's neural network is taught on the appropriate teaching selection and during the test is checked on the corresponding test selection and consequently, a classification error is calculated for all pairs. This error is equal to the part of observations in the teaching selection, the type of which was not properly defined by the neural network.

Thus, the quality of each agent for a generation is characterized by the total classification error. At the end of each generation the best agent is selected, i.e. an agent with a minimum classification error, generating the next generation of agents.

Figure 5 shows the dependence of the total classification error on the number of generations for the best agent and the average for the population. It is evident that by the 20th generation the error is reduced to 3%. The best agent's error is 1%.

Multiagent system structure [5, 6], in which agents are represented as nodes in the network is shown in Fig. 6.

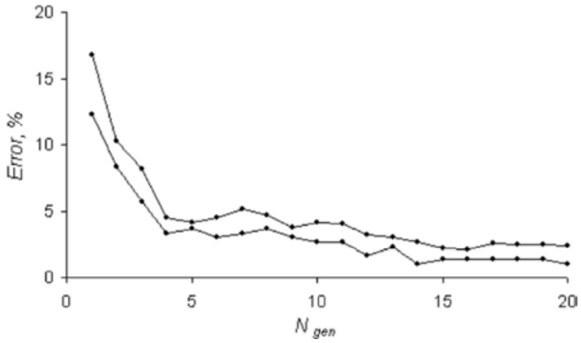

Fig. 5: Dependence of the total classification error on the number of generations. The upper curve - average for the population, the lower - for the best agent

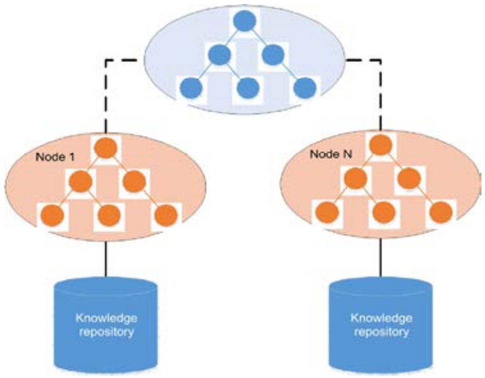

Fig. 6: Multiagent system structure

Images were captured by the video camera (web-camera), resolution 640x480, 8-bit, 30 frame/sec. Alphabet of recognizable gestures consisted of ten gestures: Letter Z, Letter M, Letter P, Letter N, Letter W, Wave, Infinity, Triangle, Square and Cross. This alphabet was selected based on considerations of application of gestures not used in normal communication and consisting of the basic gestures of sign language, intuitive for a user.

M matrix transfer to the remote server was carried out at a frequency of 15 Hz (every second frame).

Estimation of the Data Transfer Volume in the Matrix Form: The most convenient and commonly used methods of data presenting for transfer via http protocol are XML – eXtensible Markup Language and JSON – JavaScriptObjectNotation (text data interchange format).

For normal recognition operation, a minimum of 10 frames per second should be transferred, respectively, equal to 10 matrices per second.

Mâ matrix in XML Θ JSON can be presented as follows [7, 8]

XML:

```
<?xmlversion="1.0"encoding="utf-8"?>
<matrix>
<row>
<cell>0,98</cell>
<cell>0</cell>
<cell>-0.17</cell>
<cell>0</cell>
</row>
<row>
<cell>-0.06</cell>
<cell>0.94</cell>
<cell>0.34</cell>
<cell>0</cell>
</row>
<row>
<cell>0.16</cell>
<cell>-0.34</cell>
<cell>0.93</cell>
<cell>0.8</cell>
</row>
<row>
<cell>0</cell>
<cell>0</cell>
<cell>0</cell>
<cell>1</cell>
</row>
</matrix>
```

JSON:

```
matrix = [
[0.98,0,-0.17,0],
[-0.06,0.94,0.34,0],
[0.16,-0.34,0.93,0.8],
[0,0,0,1]
]
```

Order of matrix in xml format is approximately 500 bytes, in json format it is 100 bytes. Thus, for 1 second transmission 500×10 = 5000 bytes must pass for xml format and 10×100 = 1000 bytes for json. Therefore, it is advisable to use json format. In addition, there are libraries in various programming languages that allow the objects to be serialized in and deserialized out of this format, which can simplify the data exchange programming process.

Data Transmission Rate in a Distributed System: The most common wireless data networks are IEEE 802.11 or Wi-Fi, as well as the data network through mobile

Table 1: Capacities of most common data channels

Protocol	Ideal Tx/Rx rate	Actual Tx/Rx rate
GPRS	up to 171.4 Kb/s	48/24 Kb/s
EDGE	up to 474.6 Kb/s	
UTMS	up to 384 Kb/s	
HSPA	3.6 / 2 Mbit/s	2 / 0.3 Mbit/s
WiMAX	up to 40 Mbit/s	5-6 Mbit/s
Wi-Fi	up to 150 Mbit/s depending on the protocol version	20-24 Mbit/s
LTE	100 Mbit/s / 50 Mbit/s (20 Mbit/s / 10 Mbit/s is limited by operators)	up to 10 / up to 5 Mbit/s

operators. Wireless networks are most preferred for the average user. In the latter case, the operators provide the opportunity to transfer data via the protocols intermediate between 2G and 3G, such as GPRS and EDGE (EGPRS); 3G standards – UTMS, HSPA. Also, large operators carry out testing or implementing LTE networks, which are the fourth generation networks or 4G. Also, the transition standard from 3G to 4G, namely WiMAX, which was promoted by Yota, is worth mentioning. However, at this time the company actively promotes LTE standard.

Consider the capacities of most common data channels at the moment. Comparative characteristics can be represented as Table 1.

Unfortunately, the ideal rate is very rarely attainable due to a variety of external factors and a variety of possible hardware configurations.

For the video stream processing of the input gesture recognition commands, a number of different ways may be implemented. They differ in data transmission volume and processing rate. There are 3 main ways to transfer data:

- Streaming video directly from the client to the server.
- Head movement information transfer in the form of a vector from each frame
- Data transfer with the characteristics of already recognized gestures.

When transmitting streaming video the capture is done with web-camera or mobile phone camera, it is immediately, without pre-processing, sent to a server where all the necessary manipulations to determine the gesture are performed. Compression and packaging in mp4 format is not considered pretreatment as in most cases it is supported by built-in means.

The size of one second of video resolution of 640×480 and 29 fps (frames per second) in mp4 format without audio recording is approximately equal to Size = 350 KB (2800 Kbit or 2.8 Mbit). Measurements were carried out using the Samsung Galaxy S smartphone camera. Thus, one 640×480 frame in the video stream has a Size / fps = 12 KB.

Using mobile devices before saving the video to a memory card or phone memory will not allow obtaining the desired number of frames per second. The real FPS value varies from 7 to 15. Thus the size of a second of streaming video from a mobile device will be in the range from 84 KB (672Kb) to 180KB (1440 Kb).

Not every channel is suitable for transmission of the amount of video with the quality as claimed, to the server in real time or with minimal delay. WiMax would be the minimum required of the above-mentioned methods of commands' transmitting and processing.

Regarding the foregoing it may be concluded that this method is suitable for constant connection via WiMax and faster channel. Another advantage of this method might be the low demand to the computing power of the device, because all the work to extract gestures will be performed on the server.

Consequently, at low throughput of Internet channel the matrices would be preferably operated and transferred to the server for processing in json format.

CONCLUSION

The proposed system for head gesture recognition forms a unified research and design information space of various systems. Due to the independence of the agents, the system can easily be expanded by adding new agents and can be the basis of promising human-computer interfaces. Since the agents that form the basis of this system are located on servers connected to the Internet/Intranet, it can help to provide support to the development teams located in different parts of the world. Each agent integrates different knowledge about the object class it represents and this leads to further integration and systematization of scientific and technical knowledge. The results presented in the article are applied in the SmartInterface system of contactless control.

REFERENCES

1. Stasevich, V.P., 2004. New principle of control systems self-learning. In the book: Extreme Robotics. SPb: SPbSPU.
2. Kotenko, I., 2002. Multiagent technology to provide intrusion detection in computer networks. In the book: Abstracts of X Russian Scientific and Technical Conference "Information Security Methods and Technical Means." SPb: SPbSPU, 2002.

3. Luger, D.F., 2003. Artificial Intelligence: Strategies and methods for solving complex problems. – M.: Publishing House “Williams,”
4. Narushev Ye. S. and V.F. Khoroshevsky, 2000. AgSDK: multiagent systems development tools. // Proceedings of the conference CAI-2000. M.: FML,
5. Guestrin C., M. Lagoudakis, et al., 2002. Coordinated reinforcement learning. In Proc. of the 19th Int. Conf. on Machine Learning, Sydney, 2002.
6. Stone, P. and M. Veloso, 2000. Multiagent systems: a survey from a machine learning perspective. Autonomous Robots, 8(3).
7. Morozova T. Yu, M.A. Chistyakova and D.A. Akimov, 2013. Revisited creation of contactless means of managing large data amounts in ergatic systems. Aerospace Instrumentation. 5: 46-56.
8. Morozova, T. Yu. and D.A. Akimov, 2013. Use of active external appearance models when designing contactless interface for mechatronic systems control. Devices and systems. Management, monitoring, Diagnostics. 1: 2-12.
9. Ostroukh, A.V., M.N. Krasnyansky, T.L. Davydova, O.O. Varlamov, 2011. Analysis of possibilities of using mivar technology in systems of artificial intellect and modern robotics. Transactions TSTU. 17(3): 687-694.
10. Ostroukh, A.V., R.A. Sandu and O.O. Varlamov, 2011. Mivar automated control systems of technological processes for oil industry of Russia. Automation, telemechanization and communication in the oil Industry. 11: 37-41.
11. Ostroukh, A.V., 2008. Bases of construction of artificial intelligence systems for industrial and construction enterprises. Moscow: «Tech Poligraph Center», 280 p. ISBN 978-5-94385-033-2.
12. Ostroukh, A.V., S.A. Vasuhova and M.N. Krasnyanskiy, 2011. A. Samaratunga. Study the prospects and problems of integration of human-computer: artificial intelligence, robotics, technological singularity and virtual reality // Prospects of Science. 4(19): 109-112.
13. Vasuhova, S.A., O.I. Raja Ram Chaudhary and Maksimychev, A.M. Vas'kovsky, 2012. Modeling the behavior of intelligent robot // In the world of scientific discoveries. 2.6(26): 110 - 114.
14. Ostroukh, A.V., 2012. Input and processing of digital information. Moscow: Publishing House "Academy". 2012. 288 p. - ISBN 978-5-7695-9457-1.
15. Ostroukh, A.V., 2013. Artificial intelligence systems in industry, robotics and the transport sphere. Krasnoyarsk: Research and Innovation Center, 2013. 326 p. - ISBN 978-5-906314-10-9.