Methods of Primary Processing Handwriting Samples at User Authentication Using a Probabilistic Neural Network

Anatolii Davydenko[1[0000-0001-6466-1690]], Olena Vysotska[2[0000-0002-9543-1385]] and Tetiana Shmelova [2[0000-0002-9737-6906]]

[1]Pukhov Institute for Modeling in Energy Engineering of NAS of Ukraine, Kyiv, Ukraine
davidenkoan@gmail.com

[2]National Aviation University, Kyiv, Ukraine
Lek_Vys@ukr.net

Abstract. This article analyzes the dynamic biometric methods to authenticate users of automated systems. We have feasibility of their use for the organization of access and privacy of information in automated systems. For further analysis and use authentication methods are selected keystroke pattern and handwriting. For each of these methods, many handwriting features are created to analyze them during user authentication. As the mechanism of recognition selected probabilistic neural network as a type of neural network suitable for solving the problem of object recognition. After that, the proposed technology primary processing of handwriting samples through which achieved increase the probability of correct recognition of users. The stages of the implementation process considered and their expediency is proved. It is formulated what and why errors may occur in samples of handwriting, which ones should be removed or which ones should be corrected. For the method of authenticating users by handwriting, the technology of selecting the most significant points, whose characteristics it is advisable to analyze during recognition, is proposed. The quality criterion of the analyzed characteristics is also proposed. Further, a number of experiments were performed using the developed software. The results of these experiments prove the correctness of the proposed methods and the effectiveness of the proposed technology of primary processing of handwriting samples.

Keywords: multifactor authentication, identification, authentication, biometrics, keystroke pattern, handwriting, primary processing handwriting samples.

1 Introduction

In recent years, computer technology is probably already involved in all areas of activity. All information is stored and processed using either the existing software or software what specifically developed for the implementation of business process typical organization. An important aspect is to ensure the confidentiality of information and implementation of access control. The function of user

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CybHyg-2019: International Workshop on Cyber Hygiene, Kyiv, Ukraine, November 30, 2019

CEUR-WS.org/Vol-2654/paper56.pdf

authentication is a critical component in the organization of access and privacy of information in the software used, so it is urgent to develop scientific problem.

There are various methods of authentication, including authentication using keys, digital signatures, passwords, biometrics, etc. [1-8]. Where the level of information security put forward higher requirements, it makes sense to use multifactor authentication. This paper uses biometric authentication methods. The advantages of biometric authentication from other methods of solving this problem is the fact that: a person is a carrier of their "biometric password", ie, no need to remember the password and can not be somewhere to forget or lose; high degree of unique passwords; difficult to falsify a password. And the use of methods for recognizing methods based on the analysis of dynamic biometric characteristics, such as keystroke pattern and handwriting, has additional advantages, namely: they do not require additional expensive equipment (for keystroke pattern); allows to increase the degree of multifactor authentication of users of information systems; can be used not only to authenticate users but also for monitoring their work. This approach provides significant advantages for businesses that critically depend on their workers' level of attention during work. In case, if monitoring fixes, a significant deviation of the characteristics of the user's handwriting from their average statistical values for this user, thereby fixing the anomalous state of users for some reason (external factors, illness, etc.) distracted from work, or unauthorized change of the user. All of it makes relevant and expedient the use of biometric authentication methods and especially dynamic biometric methods, among which are the recognition of users by their keystroke pattern and handwriting.

But with all these advantages, biometric recognition methods based on dynamic characteristics analysis have some drawbacks. The main one is that a person's dynamic characteristics are characterized by some instability. In addition, some instability in the characteristics of handwriting is caused by the features of the devices used to transmit a sample of handwriting to the computer. In this work, a probabilistic neural network is chosen to solve the problem of user recognition [2-4,9]. This kind of neural networks [9-10] copes with the problem of pattern recognition well enough, but only if there is a slight instability of the analyzed characteristics in the sample, but more significant deviations need to be corrected or removed from the recognition process to ensure high probability proper recognition. That is why this article is dedicated to the primary processing of keystroke pattern and handwriting samples. That is, the task of this work was to develop methods for the primary processing of samples of keystroke pattern and handwriting to increase the probability of correct recognition of users.

2 Problem statement

This work improves the technology of authentication of users of automated systems by their keystroke pattern and handwriting. The operation of any biometric authentication system consists of the following two modes [2-4]:

1. Registration of users in the automated system, that is, the creation of multiple legal users of this system $\boldsymbol{USL}=\{\bigcup_{t=1}^{l} USL_t\} = \{USL_1, USL_2, \ldots, USL_t, \ldots, USL_d, \ldots, USL_l\}$, where $t=\overline{1,l}$; l – number of legal users; USL_d – a legal user who appears to be an authorized, authenticated party. In this mode, there is an accumulation of a database of training samples of the biometric characteristics of the user being analyzed. That is, many training samples are created in the training samples database $\boldsymbol{O}=\{\bigcup_{t=1}^{l}\bigcup_{j=1}^{z_t} \boldsymbol{O}_{t,j}\}$; $\boldsymbol{O}_{t,j}=\{\bigcup_{i=1}^{ks\_k} u_{t,j,i}\}$, where, respectively, the set $\boldsymbol{O}$ takes the following form: $\boldsymbol{O}=\{\bigcup_{t=1}^{l}\bigcup_{j=1}^{z_t}\bigcup_{i=1}^{ks\_k} u_{t,j,i}\}$; $j=\overline{1,z_t}$; z_t – the number of training samples t-th use in the training samples database; u_{ijt} - the value of the i-th characteristic in the j-th training sample of the t-th user; ks_k - is the number of characters in the keyword dynamics input or writing, which is analyzed in the authentication. In this case, samples of keystroke pattern and handwriting are respectively accumulated. The volume of this database depends on what biometric characteristics of the person are being analyzed. For dynamic performance, it is desirable to accumulate at least several hundred samples for each user. It is also advisable to update this database periodically.

2. Authentication of registered users. In this mode, the user who undergoes the authentication procedure presents his biometric password, the automated system uses authentication mechanism, in this case probabilistic neural network, authenticates by comparing the presented sample with the samples stored in the training samples database. As a result of the recognition, the neural network determines the username to which the user who entered the biometric password most likely belongs. If this name is the same as the user name, then the authentication process is considered successful.

As stated earlier, dynamic biometric characteristics are characterized by some instability that reduces the probability of correct user recognition (P) when used. To correct this shortcoming, this work develops a technology of primary processing of handwriting samples, which corrects some errors and removes them, thereby increasing the correct recognition of users. The effectiveness of this technology in this work is tested using experiments.

3 The solution to this problem

In order to solve this problem, a database of keystroke pattern and handwriting training samples was first accumulated, respectively. The list of features that make up a handwriting sample depends on the device used to send the handwriting sample to your computer. A standard keyboard is used to convey a sample of keystroke pattern. In this work, a graphic tablet is used to convey a sample of handwriting as one of the touch screen devices. When authenticating by keystroke pattern, the user is prompted to enter some kind of password (key phrase), the dynamics of the input

of which is analyzed for recognition. When authenticating by handwriting, the user is asked to write the word-password on the graphic tablet, then the word that is written is analyzed for correctness and the dynamics of spelling of the word. At the same time, users are conditioned that all characters of the key phrase should be written not separately, but separately. This condition greatly simplifies the recognition process and reduces the resources involved in this biometric authentication systemIn both cases some Ukrainian word, word-combination or letter combination is used as a keyphrase. When recognizing **by keystroke pattern, the following features can be analyzed** for each word-password character: the value of the time intervals between entering two adjacent password characters; the value of time intervals between the release of two adjacent characters; the value of the time intervals between pressing and releasing each password character. You can analyze all of these features, or some of them, at the same time. In addition, in this work, each sample handwriting stores the number of keys that were mistakenly pressed when creating the sample. In **handwriting recognition, the following features can be analyzed** for each point in the word-password symbol: the value of the X coordinate; the value of the coordinate Y; point type; the pressure value at which the user presses a handle (or other similar device) on the touch screen when creating points; the value of the angle of change of direction of writing when creating points; the amount of time that elapsed from the beginning of the character to the specific point; the value of the speed of movement of the handle from the forward point to a given point; as well as for images of each word-password character: the value of the symbol image area; the value of the number of points of the symbols to be analyzed during recognition; the value of the angle of inclination of the characters; the value of the frequency of the symbol points, which is fixed by the system; the value of the number of repetitions of points in the symbol plot (points created in succession with the same two coordinates) [2-4].

Some stages of initial processing are required for samples of both handwriting and keystroke pattern, but most of the steps are inherent in certain biometric characteristics. Next, let's look at the common processing steps and the processing required for handwriting and keystroke pattern samples [2-4].

First, you need to choose the correct word, password, keyboard input or typing using the graphical tablet being analyzed. It is advisable to choose as a key phrase the text that is often used in the field in which the organization that uses this biometric authentication system operates. If a person often types or writes specific words or phrases, then he develops a characteristic handwriting, which makes sense to analyze for recognition of that person.

Second, not all typing or typing characteristics have the same quality of recognition for a particular group of users. That is, in each specific organization, for each group of people it is necessary to choose those features of handwriting that are most characteristic.

Third, there will always be error samples in the training sample database, which is required. Some of these errors are specific to the individual user, so they should be left in the database, but if the deviation is significant, then using this sample handwriting as a training will have a negative impact on the probability of correct

recognition. For example, if a certain user pauses a lot before typing a particular character than before typing other keys, or writes a larger character than other letters, then the characteristic features of that user's handwriting are not correct. But if a person is distracted or sick and as a result, the speed of typing is significantly reduced or the number of misspelled keys is too high, or when writing a password on the graphic tablet besides the desired text, the user wrote something superfluous, then this sample is wrong and it should be removed from the database of training samples. In addition, there are errors that are associated with the specific use of equipment needed to transmit handwriting to your computer. For example, if a user wrote the password in the upper left corner of the tablet today, and writes 2 inches below tomorrow, then without using some data correction, these patterns should be perceived by the system as samples of different users' handwriting or patterns of different passwords.

That is, to summarize all of the above, then the primary processing of handwriting samples should include the following steps [2-4]:

1. Choosing the right password word (key phrase) whose typing or writing is specific to a specific group of people of a particular organization.

2. Selection of handwriting characteristics that should be analyzed when entering (writing) this key phrase by user data.

3. Deletion or correction of handwriting samples containing erroneous data.

Next, let's take a closer look at the criteria for selecting the optimal keyword phrase and the attributes of its input (writing) that will be analyzed for recognition. Then let's look at what data is wrong for each of these types of handwriting and which of these errors should be deleted and which ones should be corrected. In addition, consider the algorithm for performing these actions.

In order to select the optimal keyword phrase and the characteristics of its input (spelling) that will be analyzed for recognition, it is necessary to first accumulate a database of small volume training data (trial database) for different variants of the word password. Then choose the best keyword phrase and then accumulate a complete database of training samples of the required volume for the selected keyword phrase.

There are different techniques for determining the best handwriting for recognition. In this paper, it is proposed to use the H_i, function, which is calculated on the basis of characteristics such as mathematical expectation M_{it} (1) and dispersion S_{it} (2) for each trait being analyzed.

$$M_{it} = \frac{\sum_{j=1}^{z_t} u_{ijt}}{z_t}; \quad (1)$$

$$S_{it} = \sqrt{\sum_{j=1}^{z_t} (u_{ijt} - M_{it})^2 / (z_t - 1)}. \quad (2)$$

Accordingly, the function H_i, is calculated by the formula (3):

$$H_i = \frac{\sum_{r=1}^{l-1} \sum_{c=r+1}^{l} \frac{(S_{ir} + S_{ic})}{|M_{ir} - M_{ic}|}}{komb}, \quad (3)$$

where *komb* - is the number of possible combinations from *l* to two (the number of possible pairs for comparison), which are calculated by the formula (4)

$$komb = \frac{l!}{2!*(l-2)!} = \frac{l!}{2*(l-2)!}. \quad (4)$$

The main requirement for the traits being analyzed is that H_i<1, but the smaller the H_i value, the better the recognition for the trait. As a definition of H_i for all characteristics, it is necessary to select the key phrase in which the highest number of features with the best H_i value is and in this phrase to select the desired number of the best features. After that, it is already necessary to accumulate a complete database of training samples.

Then, as stated earlier, erroneous samples must be removed from the accumulated database. Keystroke pattern samples and handwriting swatches have different errors.

For samples of keystroke pattern, **an error** will be called either entering the wrong word-character (first type error) or too long before performing a specific action (pressing or releasing a key) (second type error).

To determine the percentage of misspelled keys, that is, the percentage of errors of the first type, uses the formula (5):

$$Och = \frac{100 * k_o}{kol + k_o}, \quad (5)$$

where k_o - is an amount of pressures of the erroneous keys; *kol* - it is an amount of pressures of the correct keys.

In addition, the probability of correctly recognizing all users is significantly influenced by the amplitude (distribution) of the percentage of errors among users (6). The smaller the amplitude, the lower the recognition quality.

$$A_{Och} = Och_{\max} - Och_{\min}, \quad (6)$$

where $Och_{\max}$ та $Och_{\min}$ - respectively maximum and minimum error rate among all users.

To determine samples that have too long a time interval before performing a certain action, that is, samples with a second type of error, we must first calculate by the formula (7) the arithmetic mean of the *i*-th characteristic for the *t*-th user among all training samples:

$$Sr_{it} = \frac{\sum_{j=1}^{z_t} u_{ijt}}{z_t}. \quad (7)$$

Then, check each trait in the training sample on condition (8). If at least one condition is not met, then this sample must be acknowledged as a mistake of the second type of the keystroke pattern sample and deleted from the database.

$$\left|u_{ijt} - Sr_{it}\right| \le k * Sr_{it}, \quad (8)$$

where k - is the coefficient that is selected depending on the required recognition accuracy. The smaller this factor, the more accurate the training data will be accumulated, but the less it will remain after selection.

For handwriting samples, errors are in most cases related to the specificity of using a graphic tablet. They can be divided into two groups: bugs that need to be deleted and bugs that need to be fixed.

There are five types of errors that you need to remove:

1. Type 1 error is a sequence of points with zero pressure (except for the first such point in each sequence). These errors occur when the touchscreen slides its handle over the work area of the graphic tablet a short distance away, without touching it.

2. Type 2 error is a random point (small number). Occurs if the user accidentally touches a pen on the work area of a graphics tablet.

3. Type 3 error is repetition, that is, a succession of consecutive points in which the coordinate values on both axes (*X* and *Y*) have not changed (except when one of the points has zero pressure). Occurs if the data packet was transmitted to the computer due to a change of coordinates not on one of the axes (*X* and *Y*) but another parameter.

4. Type 4 error is an accidental big loss (usually with an acute angle) at the beginning of the lines. They are caused either by the inertia of the tablet or by the shaking of the user's hand.

5. Type 5 error is a poor quality pattern that is rejected due to the inability to split the key phrase image into a given number of character images. Occur if either the wrong key phrase is entered, or if the user has little experience with the graphics tablet, or if the user has written some characters not separately but together. With this error, the image of a written phrase cannot be divided into a given number of images of individual characters, so such samples are omitted.

Some of the types of errors are illustrated in Figure 1.

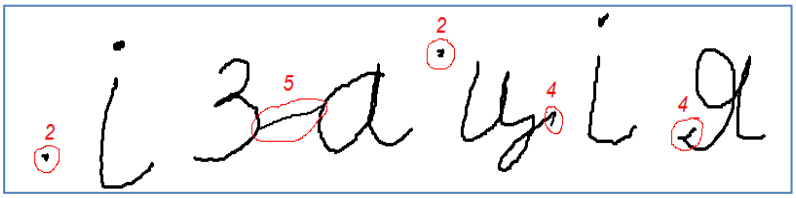

Fig. 1. Examples of erroneous data that require deletion

There are also three types of errors that need to be corrected (Fig. 2):

1. Type 6 error is a different angle, relative to the axes of the work area of the tablet, the image of a key phrase in different users. To correct such errors, a correction of type 1 is performed - character-by-character rotation of symbol images to normalize the angle of inclination of their coordinate axes (Fig. 3).

2. Type 7 error is a different location, in the work area of the tablet, the image of a key phrase for different users. To correct such errors, a correction of type 2 is performed - a character shift of images of each character to the center of the workspace of the selected size (Fig. 4).

3. Type 8 error is a different image size of a keyword phrase, on the workspace of a graphic tablet, for different users. To correct such errors, a correction of type 3 is called - a symbolic proportional mass-plotting (stretching / contraction) of images of each symbol over the entire working area of the selected size (Fig. 5).

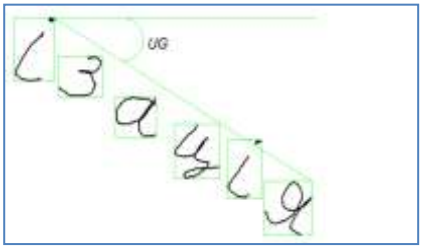

Fig. 2. Sample image before correction

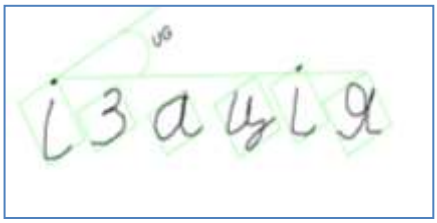

Fig. 3. The result of character rotation of the image

Removing and correcting these errors significantly increases the probability of correct recognition, which has been verified through experiments.

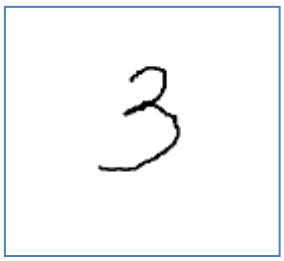

Fig. 4. The result of character shifting of the image

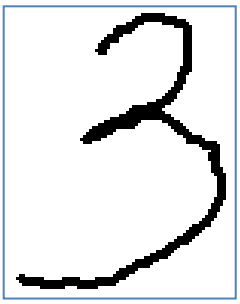

Fig. 5. Result of character-proportional scaling of the image

In addition to the already mentioned primary treatment, in this work is another very important and appropriate stage. In this paper, the characteristics of not all pixels of the word password character are analyzed, but only the most significant control points [3-4]. The necessity of this step is explained by the fact that, as a rule, the image of a single character consists of approximately 100 points, and for each character several characteristics are analyzed, and therefore too large resources are expended to perform recognition, which is not is appropriate. In this paper, there are three types of control points (Fig. 6):

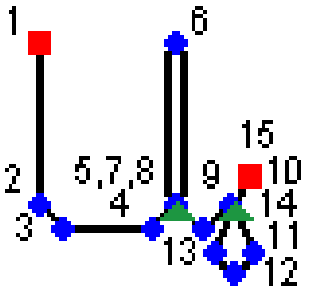

Fig. 6. Example of arrangement of control points

1. The starting and ending points of each line are the points of contact of the pen of the graphic tablet and the point of detachment of the pen from the graphic tablet. In Fig. 6 these points are indicated by squares (points 1 and 15). Stored zero data packets are used to determine such points.

2. The angular points of the lines are the points that are on the bend of the line. In Fig. 6 these points are shown by collags (points 2,3,4,5,6,7,9,10,11,12,13). The

bend of a line is called the change of direction of a line, which can be determined by the change of the sign of change of coordinates along one of the axes (or both).

3. The points of intersection are the points that are at the intersection of the lines. In Fig. 6 these points are shown by triangles (points 8 and 14). The search is complicated by the fact that the image points are not some distance from each other (the distance between the dots, beep, more than 1 pixel)

In addition, in order to use a probabilistic neural network as a recognition mechanism, all samples must have the same number of features, and due to the correct placement of control points, approximately the same number of points whose characteristics are analyzed for the images of different characters, as opposed to when the characteristics are analyzed all points of the image.

4 Experimental results

In order to verify the correct operation of the proposed methods and the efficiency of using the proposed technology of the primary work of the samples, to use the developed software, a number of experiments were conducted. In these experiments modeled the situation of implementation of biometric system authentication of users of the automated system by handwriting and keystroke pattern, respectively. In the case of keyboard handwriting, a group of 10 people were first asked to enter one of the words 200 times for each of the 3 proposed word sets. Then, using the technology indicated in the paper, a set of words ending in a combination of letters "ізація" was selected for further use. After that, each user has already been asked to enter 1,500 times one of the words ending in a combination of letters "ізація". Based on the collected data, a series of experiments were conducted in what the impact of the most critical parameters and of the implementation of the proposed technology of primary processing of handwriting samples on the probability of correct recognition were determined. During the experiments, the situation was simulated with 2 to 10 users working in the automated system. These experiments were conducted for about two weeks. In the case of handwriting recognition, the experimental conditions were about the same as those for keystroke pattern, except that users did not write the entire word on the graphic tablet, but only the letter combination ("ізація " was selected). At the same time, users were given the condition that all characters should not be written together but separately. In addition, for handwriting, for further analysis, it was determined distribution of the number of control points in the images of different characters that were formed by the technology proposed in the work. The main results of the experiments are shown in the graphs (Fig. 7-11). Consider the conditions of the experiments in more detail.

1. In the first experiment, the hypothesis of the influence of the percentage of errors on system users on the probability of their correct recognition by keyboard handwriting (*P*) was tested. This experiment was performed with the analysis of 6 handwriting characteristics, the number of training samples for each user is 1500. The results of experiments (Fig. 7) prove that the smaller the percentage of errors (*Och*), the greater the probability of correct user recognition (*P*).

2. In the second experiment, the hypothesis of the effect of the deletion of training samples with gross errors from the training sample database on the quality of the analyzed traits (H_i) was tested. This experiment was performed to authenticate users to their keyboard handwriting. The results of the experiments (Fig. 8) prove that after the removal of training samples with gross errors, the quality of the features is significantly improved, ie the H_i criterion decreases and the H_i criterion value becomes almost the same in all traits.

3. In the third experiment, the hypothesis of the effect of deletion of training samples with gross errors from the training sample database on the probability of correct recognition by keyboard handwriting (P) was tested. This experiment was performed under the condition that 6 handwriting characteristics are analyzed, the number of training samples for each user is equal to 100. The results of experiments (Fig. 9) prove that after removing the training samples with gross errors, the probability of correct recognition (P) is significantly increased, but what the smaller the deviation of at least one of the signs from its mean (k) is allowed, the greater the probability of correct recognition (P) is achieved.

4. The fourth experiment determines the distribution of values of the number of control points in the image of different characters, with the authentication of handwriting. The results of the experiments (Fig. 10) prove that at the arrangement of control points, according to the technology specified in the work, approximately the same number of points whose characteristics are analyzed is reached for images of different symbols, which makes it possible to use the probabilistic neural network as a recognition mechanism and reduces the we have the resources involved.

5. In the first experiment, the hypothesis of the effect of performing data correction (error correction of 6-8 types) in the authentication of users on their handwriting on the probability of correct recognition of the character and word-password was tested. The results of the experiments (Fig. 11) prove that the probability of correct recognition is significantly increased when performing the specified data correction.

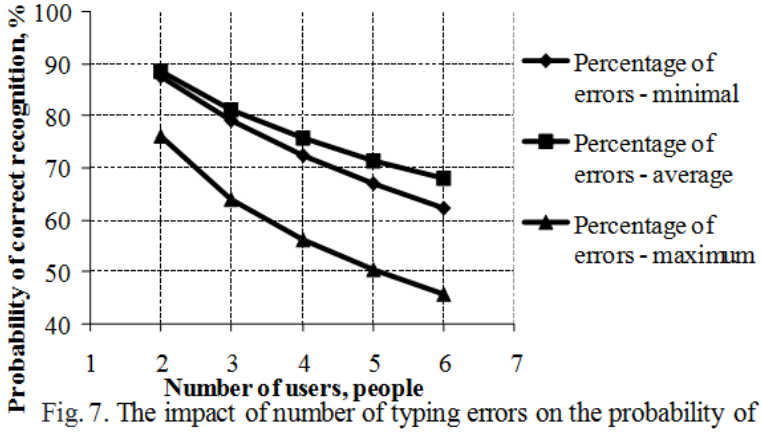

Fig. 7. The impact of number of typing errors on the probability of correct recognition of automated systems users

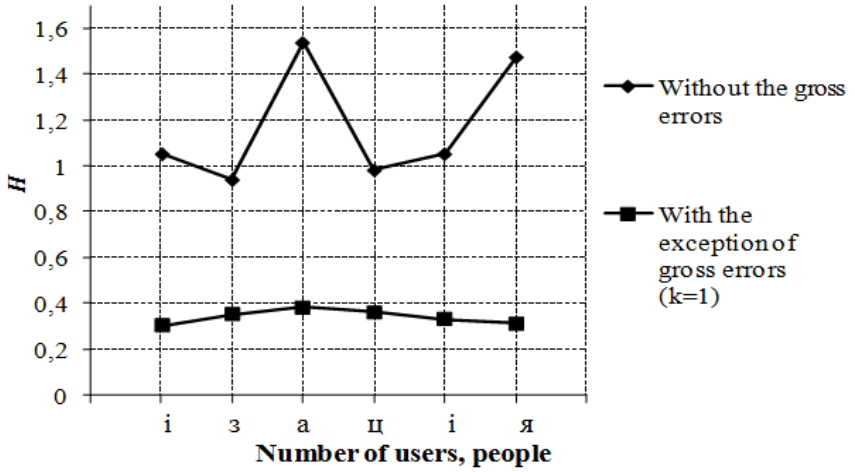

Fig. 8. Impact of exclusion of gross error on the quality of the last 6 signs in the keyword "телефонізація"

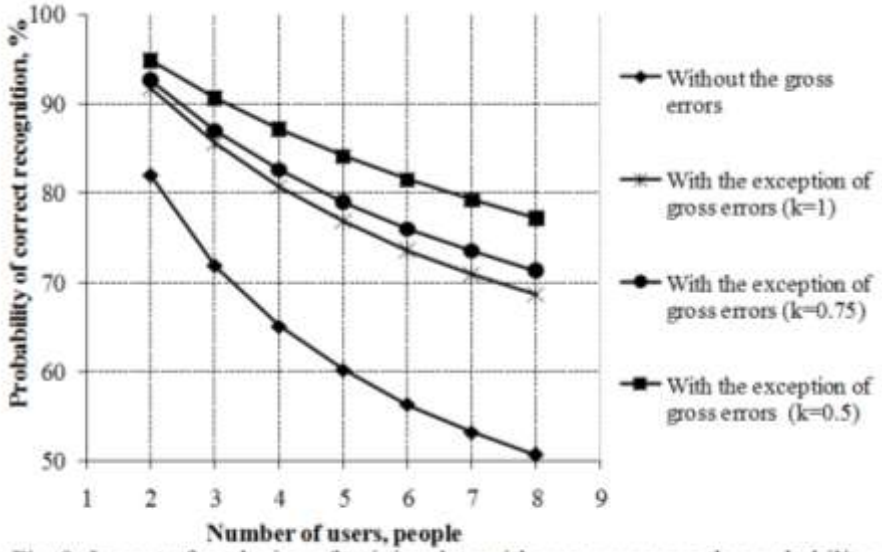

Fig. 9. Impact of exclusion of training data with gross errors on the probability of correct recognition of automated systems users

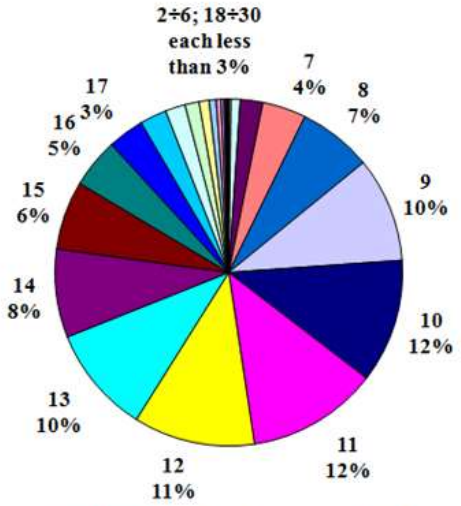

Fig. 10. The distribution of the number of control points in one character

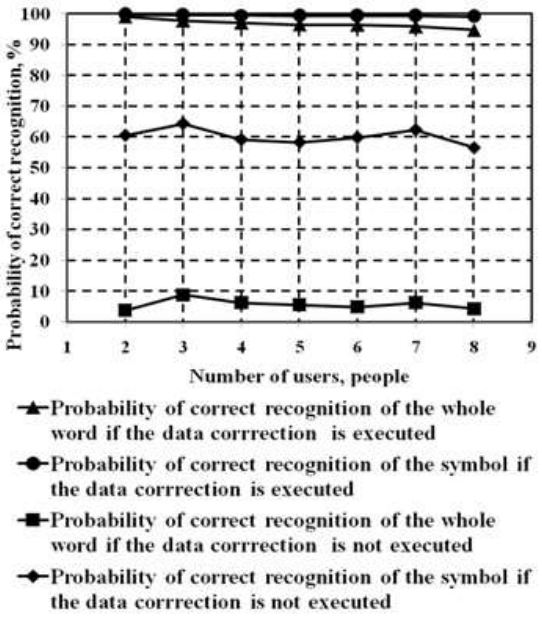

Fig. 11. Probability of correct recognition users of automated systems for their handwriting

5 Conclusions

The article is devoted to the analysis of methods of biometric authentication by dynamic characteristics. Critical features and authentication methods have been

selected to identify users by keyboard and handwriting. Primary processing of handwriting samples is proposed to reduce the neural network training term and increase the probability of correct user recognition.

References

1. Kosheva N.A., Maznychenko N.I.: Identification of users of information and computer systems: analysis and forecasting of approaches. Information Processing Systems, № 6 (113), pp. 215-223 (2013).
2. Vysotska O., Davydenko A.: Keystroke Pattern Authentication of Computer Systems Users as One of the Steps of Multifactor Authentication. Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol. 938, pp. 356-368 (2019), DOI: https://doi.org/10.1007/978-3-030-16621-2_33, last accessed: 01.11.2019.
3. Korchenko O., Davydenko A., Vysotskaya O.: Method of authentication of information systems users by their handwriting with multi-step correction of primary data. Information security, vol. 21, №1, pp. 40-51 (2019), DOI: https://doi.org/10.18372/2410-7840.21.13546, last accessed: 01.11.2019.
4. Vysotska O.O., Davydenko A.M.: Analysis of data pre-processing technology for authentication of users of computer systems by keyword pattern and handwriting. Modeling and information technologies. Collection of scientific works, vol. 55, pp. 34-41 (2010).
5. Ilyenko, A.V.: Modern ways of improving the procedure for the formation and verification of digital signature. Science-Based Technologies, 1(37), pp. 61-66 (2018), https://doi.org/10.18372/2310-5461.37.12370
6. Kazmirchuk, S., Ilyenko A., Ilyenko S.: Digital signature authentication scheme with message recovery based on the use of elliptic curves In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol. 938, pp. 279–288 (2019), https://doi.org/10.1007/978-3-030-16621-2_26, last accessed: 01.11.2019.
7. Gattal Abdeljalil, Chibani Youcef: Segmentation and Recognition Strategy of Handwritten Connected Digits Based on the Oriented Sliding Window. 2012 International Conference on Frontiers in Handwriting Recognition, pp. 297-301 (2012).
8. Furukawa Takeshi: The New Method of Identification of Handwriting Using Volumes of Indentations, 2012 International Conference on Frontiers in Handwriting Recognition, pp. 163-168 (2012).
9. Kallan R.: Basic concepts of neural networks. Translate from English. Publishing house “Williams” (2001).
10. Dychka I., Chernyshev D., Tereikovskyi I., Tereikovska L., and Pogorelov V., "Malware Detection Using Artificial Neural Networks", Advances in Computer Science for Engineering and Education II. Advances in Intelligent Systems and Computing, vol. 938, pp. 13-22 (2019).