Feature Extraction for Handwritten Signature Verification

Beresneva Anastasia, Epishkina Anna

National Nuclear Research University "MEPhI"

Abstract.

Handwritten signature verification is one of the promising methods of identity verification, as each person's signature is unique, especially if its dynamic characteristics are taken into account. The study identified more than 20 characteristics of handwritten signature. The direction of further work is devoted to the definition of significant features for the construction of an effective verification algorithm. To determine the informative features of the signature and use them in building a model of handwritten signature, it is necessary to analyze their statistical characteristics and determine how they are indicative and at the same time stable for one person.

Keywords: Handwritten signature, verification, neural network, mobile application

I. INTRODUCTION

A handwritten signature has been a hallmark for identity verification for centuries. This is a unique behavioral trait of the individual. An important advantage of signature verification compared to other biometric characteristics is its traditional use in many common commercial areas, such as legal documents, Electronic business, which includes online banking operations, financial transactions, electronic payments, access control, etc.

Figure 1 demonstrates the main stages of the online signature verification system. Currently, there are a lot of different approaches to the development of an effective verification algorithm. They differ in the stages of feature extraction and classification.

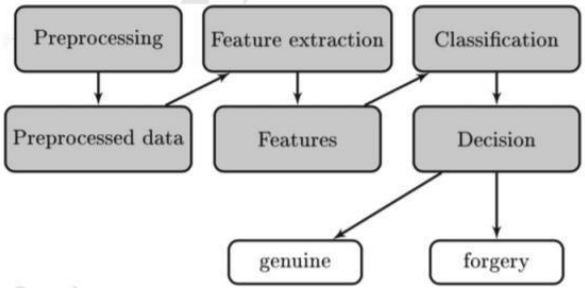


Fig. 1 Stages of a signature verification system

The effectiveness of a verification algorithm depends on the signature features it uses. The attributes must be unique and at the same time sustainable for each user. Stability is determined by the relative immutability of the characteristic from sample to sample over a long period of time. To determine stable and indicative characteristics, an experiment to collect signature samples and construct the distribution of each trait was conducted with the same set of subjects for several months.

II. SIGNATURE FEATURES

The characteristics are extracted from the individual segments of the signature into which it is broken by sharp kinks as it shown on fig.2. The study identified those informative characteristics that will uniquely determine the process of signing.

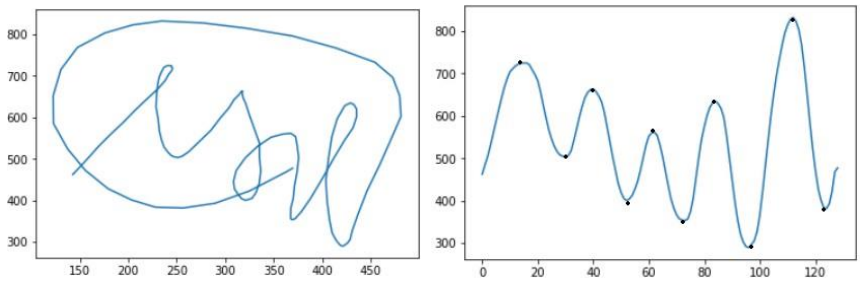


Fig. 2. Signature representation

First feature is the **number of signature segments** into which it is broken by sharp inflections in the y coordinate.
In addition to this features the analysis extracted the following:

Number of strokes of the finger up - a function that shows how many movements of the finger up made by the user. It should be noted that the last stroke does not count, since it is the end of the signature.

The aspect ratio of the signature - function, which expresses the ratio of the width (the size of the signature along the x-axis) in pixels, and normalized to pixels of the screen of your smartphone, expressed in a similar way the length of the signature (size of signature y-axis).

Signature length - a feature that represents the length of all lines drawn in the signature. The smartphone provides the user interaction data and the coordinates of the touch points of a finger. By calculating the Euclidean distance between two consecutive touch points and adding it to the current signature length, the total signature length is obtained.

Signature time - the function expresses the total time required to complete a signature. Measured in milliseconds. It is assumed that the time to commit a signature sample will always be nearly equal.

Contact time indicator - the ratio of the time during which the finger was in contact with the surface of the screen to the total time of signing. The value of the indicator is in the range from 0 to 1. For the owner of the signature, this characteristic will be permanent, since it is directly related to the time of signature.

Signature speed - the function is derived from the total length of the signature and the total time that the finger has been in direct contact with the signing area. Speed is expressed in pixels per millisecond. The trained signature should have no significant differences over time. However, this function is highly dependent on the physical and mental state of the person.

The **signature speed along the x-axis on i-th interval** is a feature representing the speed expressed in pixels per millisecond, which shows how fast the user makes the signature when considering only the xcoordinate. the Total length of the signature that the finger has passed along the x-axis is divided by the total signing time. This feature depends on the physical and mental state of the person and is used less often than other characteristics.

The **average pressure** function is achieved by monitoring the level of pressure of the finger on the screen. To get the average pressure value, it is necessary to combine all the levels of the obtained pressure into one variable and divide by the total number. The biggest influence on this characteristic is the user's body.

The **strongest pressure** moment-characteristic is expressed as a separate local signature characteristic, which can be global because it is unique to the whole signature. To extract this function, it is necessary to monitor the level of pressure that the finger exerts on the smartphone screen. The time of creation of the highest pressure level is recorded. It is assumed that the signature almost always has the same moment of strong pressure.

The **greatest slope of the device** - characteristic is expressed as local. To determine this characteristic, it is necessary to analyze the accelerometer readings for the entire signature. It is registered during the greatest inclination of the device in the process of signing.

All of the features described above are just a small subset of the informative features that can be used when authenticating users. Note that these functions are basically global characteristics of a handwritten signature and form a basic set of functions that can be used to compute some other functions, including local ones.

III. ANALYSIS OF THE DISTRIBUTION OF SIGNATURE FEATURES

During the study, a sample was formed, consisting of signatures of various users and the corresponding signatures of attackers. In order to select from all the signature features significant analysis is carried out as follows. A separate interval is considered for each user signature. At this interval, the main features that are subject to analysis are taken. Because feature values can vary for a single user from signature to signature, it is necessary to investigate the robustness of the feature. The random variable sample in this case is the value of a specific feature at a certain interval for all signatures made by the user. Next, to build a statistical distribution, the following steps were done.

A. Preprocessing

Signatures from the datasets used in the experiments are collected using an application on a mobile device written in JAVA for Android devices. The application provides a graphical interface an interface for signing and writing the obtained characteristics to a CSV file, which is then analyzed. Each participant of the experiment made 10 test signatures in order to get used to placing the signature on the screen of a mobile device and 10 signatures, the characteristics of which participated in the study.

Collected original author's signatures and signatures of fraudsters - attempts to forge the original signatures of the participants of the experiment. The result is a sample of 100 original signatures and 10 fake signatures for each, respectively.

B. Construction of a variation series.

To construct a series, it is necessary to divide the random variable into intervals, the number of which is determined by the Sturges' formula [1]:

$$k \approx \log_2 n + 1, \qquad (1)$$

where n is the sample size. The length of the interval is determined by the formula:

$$h = \frac{x^{max} - x^{min}}{m}, \qquad (2)$$

Where $R = x_{max} - x_{min}$ – the range of variation. The boundaries of the intervals are calculated

$x_0 = x_{min}, x_i = x_{i-1} + h.$

Next step is to calculate how many elements n_i falls into each i of the intervals. Then to calculate the relative frequencies at intervals as $w_i = n_i / n \; (i = 1,2,...), \hat{x}_i = (x_{i-1} + x_i)/2$.

It is also necessary to calculate the estimate of expectation and variance by formulas:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i, \quad (3)$$

$$D = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - (\hat{x})^2, \quad (4)$$

$$S = \frac{n}{n-1}D, \quad (5)$$

where $\hat{x}$ is the expectation, D is the variance, and S is the standard deviation estimate.

C. Construction of a sample of the speed characteristic

Thus, in relation to the signature features, let us consider the construction of a sample of the speed characteristic at the first interval for 35 different signatures of one user in table 1.

TABLE I: THE VALUE OF THE SIGNATURE SPEED ON THE INTERVAL.

№	v	№	v	№	v	№	v	№	v
1	0,73	8	0,77	1	0,75	2	0,76	2	0,77
2	0,75	9	0,77	1	0,76	2	0,73	3	0,70
3	0,76	1	0,76	1	0,77	2	0,74	3	0,71
4	0,69	1	0,71	1	0,78	2	0,75	3	0,72
5	0,75	1	0,72	1	0,79	2	0,76	3	0,73

6	0,8	1	0,76	2	0,71	2	0,81	3	0,75

To construct the distribution of the range of change in the value of the feature, represented as a random variable *X* in the sample volume *n*, calculate the above characteristics and enter in table 2.

TABLE II: THE ARRANGEMENT OF CHANNELS

№	i $w_i = nn\ x_{-i}$	$n_i x_{i2}$	wh_i Boundari	n es	
1	[0,68;0,70]	3	0,08	0,69	1,43	0,53
2	(0,70;0,722]	4	0,11	0,71	2	0,73
3	(0,722;0,744]	7	0,2	0,73	3,73	1,3
4	(0,744;0,766]	11	0,31	0,75	6,19	2,06
5	(0,766;0,788]	7	0,2	0,77	4,15	1,3
6	(0,788;0,81]	3	0,08	0,79	1,87	0,53
		35			19,37	

To calculate the expectation value x, calculate $x^{\frown}$ and write the results in table 2. Since $x^{\frown} = \frac{1}{n} \sum_{i=1}^{n} n_i x_i$ then in the case of velocity on the *i*-th interval we get $x^{\frown}$ = 0,747.

Next, calculate the estimate D 0,026 , to do this, enter the values $n_i x_i^2$. in table 2. After that, we calculate the value of the estimate of the mean square deviation S^2 = 0,0019.

For comparison, it is also possible to determine the estimate of the standard deviation of a random variable for a normal distribution according to the rule of three Sigma [2]. Calculate the maximum deviation from the average as $\sqrt{\square\square S_2/3\square} 0{,}13.$ The distribution is shown in Fig.3:

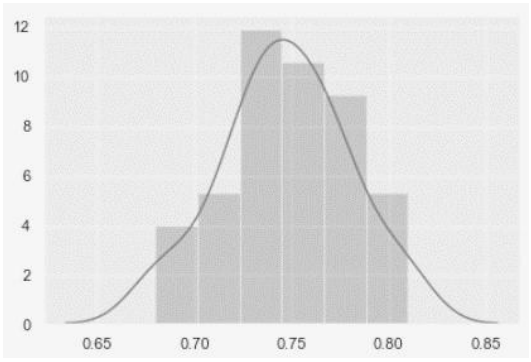
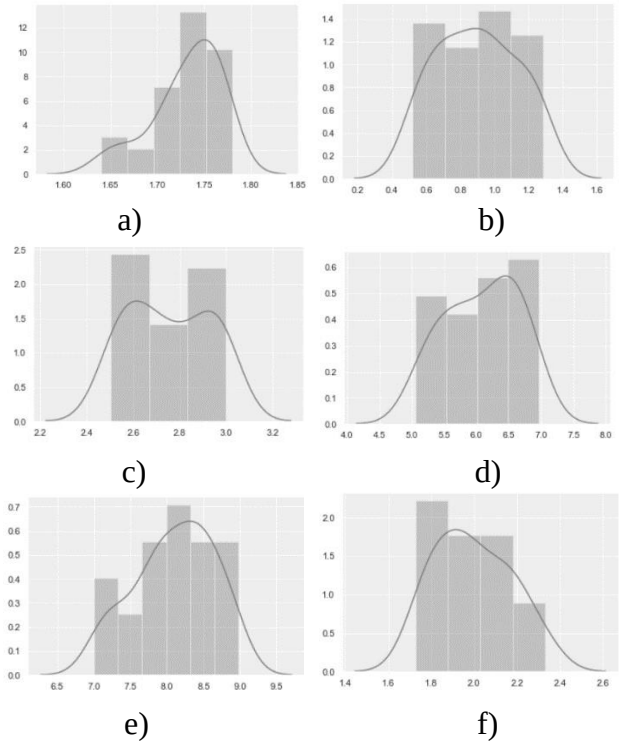
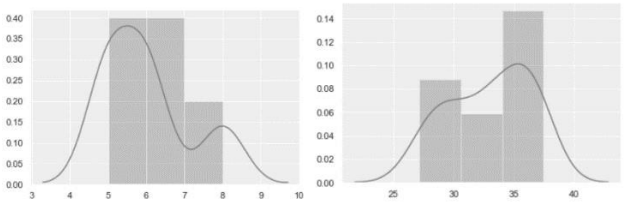


Fig. 3. Speed distribution for one user on the i-th interval.

D. Construction of distributions of signature features

Similarly, we examine each of the features extracted from the signature of one user at all signature intervals. Not all features have a normal distribution. The behavior of features that are not distributed normally cannot be analyzed with respect to the signature of the attacker, so they are not considered in the future study. For fig. 4 the distributions of all studied features are presented:





g) h)

Fig.4. Distribution of signature characteristics of one user:
a) the length of the i-th segment; b) the deviation on the i-th segment; c) the maximum forward speed;
d) the force of pressing; e) the coefficient of horizontal sweep; e) the average speed of recording;
g) the duration of the signature; h) the angle of the initial direction.

IV. STATISTICAL ANALYSIS OF SIGNATURE FEATURES

In order to identify the signature features that will identify the forgery, that is, are unique to different users, it is necessary to analyze the same features for different users and attackers relative to each other. For this purpose, we apply the mathematical apparatus of ANOVA analysis of variance (ANOVA – Analysis of Variation) [4]. This analysis is a method of checking the significance of the difference between the mean values in different samples, based on a comparison of the variances of these samples. Dividing the total variance for a single sample into multiple sources allows you to compare the variance caused by the difference between multiple samples with the variance caused by intra-group variability. The H_0 hypothesis assumes that there are no differences between the samples. If the null hypothesis is true, the estimate of variance associated with intra-group variability should be close to the estimate of inter-group variance. When false-significantly deviate.

ANOVA analysis can be applied to different samples, in our case we apply the one-factor method. In a series of experiments, we get a sample of one user's signatures, and then a sample of another user/attacker, and so on. In this case, the task is to determine whether the fact of changing the user has an impact, that is, whether this feature is distinguishable. When applying one-factor variance analysis, it is assumed that the averages of the General populations from which the samples were extracted are equal, that is, they all belong to one General population and the differences are random. To test the theories in the case of variance analysis, the F-distribution is used. F-statistics accept only positive or zero values.

The procedure of variance analysis consists in determining the ratio of intergroup variance to random variance in the measured data. As an indicator of variability, the sum of squares of deviation of parameter values from the mean is used: SS (Sum of Squares). The total sum of squares SS_{Total} is decomposed into the intergroup sum of squares SS_{BG} and the intragroup sum of squares SS_{WG}:

$$SS_{Total} = SS_{BG} + SS_{WG} \quad (6)$$

Let there be *n* samples, that is, signature samples for *n* different users. If the Ho hypothesis is correct, then both the intragroup and intergroup variances serve as estimates of the same variance

and should be approximately equal. Thus, the value of F should be close to 1 in case there are no statistically significant differences, this can be seen from the formulas (7), (8) and (9):

$$F = \frac{MS^{BG}}{MS_{WG}}, \quad (7)$$

$$MS_{BG} = \frac{SS^{BG}}{\nu BG}, \quad (8)$$

$$MS_{WG} = \frac{SS^{WG}}{\nu WG}, \quad (9)$$

The critical value of F is calculated taking into account the significance level α and the intragroup and intergroup number of degrees of freedom ν. It is quite difficult to calculate, so more often used tabular values indicating α, ν*BG*, ν*WG*. The intergroup number of degrees of freedom ν*BG* = *m*-1, where m is the number of users who signed. The intragroup number of degrees of freedom ν*WG* = *n* - *m*, where n is the number of signatures made by each user.

A. Statement of the null hypothesis

H_0 indicates the absence of differences between user signatures on the basis of speed in the *i*-th interval, in other words, all users on this basis belong to the same population and, accordingly, are equal to each other: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$.

An alternative hypothesis suggests that the signature rates of different users on the *i*-th interval are different: $\mu_1 \neq \mu_2 \neq \mu_3 \ldots \neq \mu_n$.

B. Sum of squares

In the previous step, the average value for all samples was found. Find the average value for all samples and the sum of squares $\bar{x}$ and SS_{Total}.

Then find the sum of the squares within the groups successively subtracting from each value in the group the group average SS_{WG} and the square of the deviation of each sample mean relative to the total mean SS_{BG}.

C. Fisher criteria

To estimate the statistical differences, we find the value of the Fisher criterion [5], based on the mean squares of deviations within and between groups and the corresponding degrees of freedom:

$$F > F_{cr} \text{ according } \alpha = 0{,}05 \quad (10)$$

Thus, it can be concluded that there are statistically significant differences between the groups of signature samples.

Since the value of F is greater than the critical value for a given number of observations and the number of groups, in other words, the variance between the groups contributes more to any amount of variance than that within the groups themselves, and therefore the trait is considered distinguishable.

Afterwards an analysis was performed for all features of the signature, as well as for the pairwise analysis of the original signature and the corresponding signature of the attacker. The results obtained during the study are presented in table 3.

TABLE III: SIGNIFICANT FEAURES

Feature	Distinguish ability relative to other features	Distinguish ability with respect to the attacker's signature
Speed on the *i*-th segment	+	+
Number of segments	+	+
Length of *i*-th segment	+	+
Contact time on the *i*-th segment	+	+
Maximum forward speed	-	-
Pressing force on *i*-th segment	+	-
Ratio of the horizontal scale	+	-
Average write speed	+	+
Time of effecting the signature	-	-
Angle the initial direction	-	-
Greatest slope	+	-
Stronger pressure	-	-

V. CONCLUSION

This article presents a method for determining the significant features of handwritten signatures using statistical analysis. Various static (e.g. height, slope, etc.) and dynamic (e.g. speed, pen tip pressure, etc.) signature features are extracted and analyzed to further construct an efficient verification algorithm. This method of determining the features will reduce the time of the algorithm without reducing the accuracy.

In the further research on the basis of the received features the model of the handwritten signature will be constructed. In addition, it would be useful to apply a machine learning [6] method based on the built signature model and estimate the number of errors.

REFERENCES

[1] Blanco-Gonzalo, R. Handwritten signature recognition in mobile scenarios: Performance evaluation / R. Blanco-Gonzalo, O. Miguel-Hurtado, A. Mendaza-Ormaza, R. Sanchez-Reillo. // IEEE International Carnahan Conference on Security Technology. — 2012. — P. 174-179.

[2] Jain, A. Online signature verification / A. Jain, F. Griess, S.Connell. // Pattern Recognition. — Vol. 35. — 2005. — P. 2963-2972.

[3] Zhang, K. Using landmarks to establish a point-to-point correspondence between signatures / K. Zhang, I. Pratikakis, J. Cornelis, E. Nyssen. // Pattern Analysis and Applications. — Vol. 3, no. 1. — 2000. — P. 69-75.

[4] Ao, Meng & Li, Stan. (2008). Multi-Class Classification Based on Fisher Criteria with Weighted Distance. 1 - 5. 10.1109/CCPR.2008.17

[5] Plamondon, L. Automatic signature verification and writer identification / L. Plamondon, R. Plamondon. // Pattern Recognition Journal. — Vol. 22, no. 2. — 1989. — P. 107-131.

[6] Guru, D. Online signature verification and recognition: An approach based on symbolic representation / D. Guru, H. Prakash. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — Vol. 31, no. 6. — 2009. — P. 1059-1073.