signature recognition,
normalized Levenshtein distance,
feature selection

Małgorzata PAŁYS[1], Rafał DOROZ[1], Piotr PORWIK[1]

THE USE OF METHODS OF STATISTICAL ANALYSIS IN SIGNATURE RECOGNITION SYSTEM BASED ON LEVENSHTEIN DISTANCE

The study being presented is a continuation of the previous studies that consisted in the adaptation and use of the Levenshtein method in a signature recognition process. Three methods based on the normalized Levenshtein measure were taken into consideration. The studies included an analysis and selection of appropriate signature features, on the basis of which the authenticity of a signature was verified later. A statistical apparatus was used to perform a comprehensive analysis. The independence test χ^2 was applied. It allowed determining the relationship between signature features and the error returned by the classifier.

1. INTRODUCTION

In the contemporary world is an important issue to ensure the safety of goods, resources, and data. In order to protect them, there are used common methods based on human knowledge — for example: passwords and PIN codes, as well as methods based on identifiers, e.g. identity cards and credit cards. These methods may not be able to serve their purpose for various reasons, such as forgetting a password or a PIN code, giving it to another person, or identifier loss, theft or forgery.

In the era of computerization and automation, the gap in the problems related to protections is filled by biometric techniques. One of the most popular biometric techniques is a handwritten signature. This method is widely used because of the ease of obtaining signatures, as well as due to its social and legal acceptance.

The effectiveness of the use of an analysis of handwritten signatures as a biometric technique is very high. The main factor affecting the effectiveness is selection of an appropriate signature recognition method. Currently a lot of different approaches have been proposed for signature verification in the literature [1, 3, 4, 6]. Their actions are based on different models of neural networks or Hidden Markov Models (HMM) [8]. They may also use the calculation of distance such as Euclidean or Mahalanobis [3, 6]. This study presents a method of comparing signatures with the use of the normalized Levenshtein metrics [5, 9]. The effectiveness of these metrics in the process of signature recognition has been examined [2].

The test research included selecting values of various parameters of the proposed method in order to verify signatures as accurately as possible. However, the research did not include a more detailed analysis, which would allow adequately select features of the signatures being compared. This is due to the fact that there was obtained a very large number of results, which made an analysis of the results more difficult. Such an analysis was performed under the study being presented. For this purpose, there were used statistical methods described in the study [7]. The effect of the research was determination of a combination of dynamic features of signatures, the use of which in the Levenshtein method allows obtaining the lowest error in signature recognition.

2. NORMALIZATION OF THE LEVENSHTEIN DISTANCE

The Levenshtein distance is defined as a metric for measuring the similarity of two character strings [5]. Let's define an alphabet of characters $\sum$ and a set containing all character sub-strings from this alphabet $\sum'$. Then, let's define two character strings $A=\{a_1,a_2,...,a_n\}$ and $B=\{b_1,b_2,...,b_m\}$ belonging

[1] University of Silesia, Institute of Computer Science, 41-200 Sosnowiec, Będzińska 39, Poland,
{malgorzata.palys, rafal.doroz, piotr.porwik}@us.edu.pl

to $\sum^{'}$, where *n* and *m* are the lengths of these strings. Let $T_{A,B} = T_1, T_2, ..., T_l$ mean the transformation of the *A* character string into the *B* character string with the use of the finite number *l* of elementary operations.

Elementary operations are performed on the pair of characters (a,b), where $a,b \neq \lambda$, described more often as $(a \to b)$. λ represents here an empty character, which does not belong to the alphabet. Three elementary operations can be distinguished:

- *D* – deleting a character $(a \to \lambda), (b \to \lambda)$,
- *I* – inserting a character $(\lambda \to a), (\lambda \to b)$,
- *R* – replacing a character $(a \to b), (b \to a)$.

The $T_{A,B}$ transformation can be defined for a specific path of transition from the *A* character string into the *B* character string. Let the $P_{A,B} = \{P_{A,B}^1, P_{A,B}^2, ..., P_{A,B}^h\}$ set contain all possible paths of transitions from the *A* character string into the *B* character string, where *h* is the number of all possible transition paths.

Let $W(P_{A,B})$ be a function calculating weights of individual paths from the $P_{A,B}$ set:

$$W(P_{A,B}) = \delta(T_{A,B}). \tag{1}$$

The General Levenshtein Distance (GLD) for the two character strings *A*, *B* being compared can be defined as follows:

$$GLD(A,B) = \min\{\delta(T_{A,B})\} = \min\{W(P_{A,B})\}. \tag{2}$$

As the final value of the Levenshtein distance calculated for two character strings is included in the $[0,\infty)$ interval, it is not possible on this basis to determine the percentage similarity of the strings being compared. This considerably hinders the evaluation of similarity of the strings being compared.

*Ned*1 metric is defined by the formula:

$$Ned1(A,B) = \min\left\{\frac{W(P_{A,B})}{Ld(P_{A,B})}\right\}, \tag{3}$$

where:

$Ld(P_{A,B}) = |P_{A,B}|$ – the number of elementary operations in an individual path.

Another measure is the *Ned*2 metric described by the following formula:

$$Ned2(A,B) = \min\left\{\frac{W(P_{A,B})}{|A|+|B|}\right\} = \frac{GLD(A,B)}{|A|+|B|}, \tag{4}$$

where:

$|A|+|B|$ – is the sum of lengths of the *A* and *B* strings.

Another modification of the Levenshtein distance, used in this study, is the $d_{N\text{-}GLD}$ distance. This distance is expressed by the formula:

$$d_{N-GLD}(A,B) = \frac{2 \cdot GLD(A,B)}{\max(D,I) \cdot (|A|+|B|) + GLD(A,B)}, \tag{5}$$

where:
D – the cost of deleting a character,
I – the cost of inserting a character.

All presented metrics: *Ned*1, *Ned*2, $d_{N\text{-}GLD}$ return results from the [0,1] interval. If two strings being compared are the same, the metrics return the 0 value. For further assessment of their effectiveness with the use of EER, the metrics (3), (4) and (5) were adequately modified so that the result of the comparison of two identical strings was the value 1:

$$NED1(A,B) = 1 - Ned1(A,B), \tag{6}$$

$$NED2(A,B) = 1 - Ned2(A,B), \tag{7}$$

$$NGLD(A,B) = 1 - d_{N-GLD}(A,B). \tag{8}$$

Thanks to a tablet, a signature can be recorded in the form of an n-point set [2]. Values of individual features are determined in each point. Up to now, about 40 different signature features have been identified [4]. The evaluation of the similarity of individual signatures was performed on the basis of an analysis of three signature features:

1. $X = \{x_1, x_2, ..., x_n\}$ – x coordinates of signature points,
2. $Y = \{y_1, y_2, ..., y_n\}$ – y coordinates of signature points,
3. $P = \{p_1, p_2, ..., p_n\}$ – pen pressure on the tablet surface in successive signature points.

Thus, many different values were obtained as the result of the comparison, and each of them described the similarity of a different signature feature. Then the F_i weight was assigned to each M_i value that determines the similarity of the i-th feature in two signatures being compared. This allowed determining, which of the analysed features were most important, and how considerable influence on the effectiveness of the signature recognition process they have. The formula for determining the *Sim* similarity value of two signatures S_1 and S_2, taking into account m features, is as follows:

$$Sim(S_1, S_2) = \sum_{i=1}^{m} (M_i \cdot F_i), \text{ for } F_i \in [0,1], \sum_{i=1}^{m} F_i = 1. \tag{9}$$

It has been assumed that the weights of individual signature features will change within the range from 0.0 to 1.0 with the 0.2 increment and, that the sum of the weights of all features must equal 1.0.

3. THE COURSE AND RESULTS OF THE STUDIES

The results obtained with the use of *NED*1, *NED*2 and *NGLD* measures were analysed. The studies were limited to a statistical analysis of a combination of three basic features, the values of which are sent directly from a tablet, that is *X*, *Y* and *P* features. The created combinations are *XYP*, *XY*, *XP*, *YP*. In order to assess, which of them has the greatest impact on EER values, the χ^2 test was used. It allows determining whether there is a relationship between feature combinations and EER values. Sample data obtained using the Levenshtein algorithm had the following format:

0.6_0.2_0.2_1.812,

where the first value is the F_1 weight of the *X* feature, the second is the F_2 weight of the *Y* feature, and the third is the F_3 weight of the *P* feature. The fourth value defines EER determined for the assumed values of weights of individual features. As the number of results for each of the three analysed measures was very high (18231), the analysed data were divided into 7 subsets. Each subset was assigned with a different EER range. Boundaries of division are determined by dividing the range between the highest and lowest value into 7 equal parts. Each range was named depending on the value of the errors it contained. For example, for the *NED*1 measure (in which the lowest value of EER = 1.161%, and the highest value of EER = 46.667%), the determined ranges are presented in Table 1.

Table 1. Table of ranges of EER values determined for the *NED*1 measure.

Name of range	Range EER [%]
Excellent	[1.161-7.662)
Very good	[7.662-14.163)
Good	[14.163-20.664)
Average	[20.664-27.165)
Poor	[27.165-33.666)
Bad	[33.666-40.167)
Very bad	[40.167-46.668)

Basing on the assumptions presented in Table 1, the quantity table was prepared, which contains the quantity of EER values obtained for different combinations of signature features (Table 2).

Table 2. Table showing the quantity of EER values for the *NED*1 measure.

	XYP	*XY*	*XP*	*YP*
Excellent	7029	2190	1110	1764
Very good	6574	1887	2078	1842
Good	3154	811	1338	1005
Average	1094	234	489	419
Poor	300	60	151	131
Bad	71	15	31	33
Very bad	9	3	3	1
$\sum$	18231	5200	5200	5195

In order to perform the χ^2 test, two hypotheses should be made: H_0 and H_1. The null hypothesis H_0 assumes that selection of features does not affect the effectiveness of signature comparison using the Levenshtein method:

$$H_0 : P\left(Z = z_k \cdot \ W = w_m\right) = P\left(Z = z_k\right) \cdot P\left(W = w_m\right), \tag{10}$$

where the variable Z is a combination of the X, Y, P features, while the variable W is a range of EER values.
However, the alternative hypothesis H_1 shows a relationship between the Z and W variables:

$$H_1 : P\left(Z = z_k \cdot \ W = w_m\right) \neq P\left(Z = z_k\right) \cdot P\left(W = w_m\right) \tag{11}$$

at the adopted level of significance α.
Then the expected quantities should be calculated using the formula (12):

$$E_{pt} = \frac{\sum_{j=1}^{m} n_{pj} \cdot \sum_{i=1}^{k} n_{it}}{\sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}} \tag{12}$$

where:
k – number of rows in Table 2,
m – number of columns in Table 2,
n_{ij} – an element in the intersection of row i and column j of Table 2.

Table 3 contains the expected quantities determined for the *NED*1 measure.

Table 3. Table of expected EER values for the *NED*1 measure.

	XYP	*XY*	*XP*	*YP*
Excellent	6517.693	1859.032	1859.032	1857.244
Very good	6672.915	1903.305	1903.305	1901.475
Good	3399.786	969.716	969.716	968.783
Average	1205.124	343.736	343.736	343.405
Poor	346.015	98.693	98.693	98.598
Bad	80.845	23.059	23.059	23.037
Very bad	8.623	2.460	2.460	2.457

Table 4 shows the calculated differences between actual quantities (Table 2) and expected quantities (Table 3):

Table 4. Table showing the difference between the actual quantities and expected quantities of EER values for the *NED*1 measure.

	XYP	*XY*	*XP*	*YP*
Excellent	511.307	330.968	-749.032	-93.244
Very good	-98.915	-16.305	174.695	-59.475
Good	-245.786	-158.716	368.284	36.217
Average	-111.124	-109.736	145.264	75.595
Poor	-46.015	-38.693	52.307	32.402
Bad	-9.845	-8.059	7.941	9.963
Very bad	0.377	0.540	0.540	-1.457

Having the above data, the statistic can be determined using the following formula:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{13}$$

where:

k – number of rows in Table 2,
m – number of columns in Table 2,
n_{ij} – an element in the intersection of row i and column j of Table 2,
E_{ij} – an expected quantities in the intersection of row i and column j of Table 2.

For the *NED*1 measure, the calculated statistic is χ^2 = 805.128. The critical value $\chi^2_\alpha = 28.869$ was taken from the distribution tables χ^2 for the adopted level of significance α = 0.05 and for s = (k-1)(m-1) = 18 degrees of freedom. The calculated statistic belongs to the critical area ($\chi^2 > \chi^2_\alpha$). Therefore the null hypothesis should be rejected in favour of the alternative hypothesis that assumes that these combinations affect the range of EER values. In addition, basing on Table 4, it can be stated that the greatest impact on the EER value in the Levenshtein method has a combination of *XP* features, and therefore the use of this combination will allow increasing the effectiveness of signature comparison by this method.

A similar analysis was carried out for the *NED*2 and *NGLD* measures. Statistics for the *NED*2 and *NGLD* measures are respectively $\chi^2 = 2749.321$ and $\chi^2 = 2646.392$. Thus, they belong to the same critical area as the *NED*1 measure. There was performed also an analysis of the tables showing the difference between actual quantities and expected quantities of EER values for the *NED*2 and *NGLD* measures. These tables are presented below.

Table 5. Table showing the difference between the actual quantities and expected quantities of EER values for the *NED*2 measure.

*NED*2	*XYP*	*XY*	*XP*	*YP*
Excellent	909.398	580.349	-1322.651	-167.096
Very good	-131.497	119.092	220.092	-207.688
Good	-374.777	-279.005	546.995	106.788
Average	-224.244	-244.013	329.987	138.270
Poor	-117.355	-120.917	156.083	82.190
Bad	-49.733	-46.434	58.566	37.601
Very bad	-11.792	-9.072	10.928	9.935

Table 6. Table showing the difference between the actual quantities and expected quantities of EER values for the *NGLD* measure.

NGLD	*XYP*	*XY*	*XP*	*YP*
Excellent	945.170	497.869	-1288.747	-154.293
Very good	-156.868	145.144	237.457	-225.732
Good	-377.857	-239.446	512.753	104.550
Average	-232.937	-242.276	325.794	149.419
Poor	-115.539	-109.912	149.113	76.337
Bad	-50.256	-42.612	53.397	39.471
Very bad	-11.713	-8.768	10.233	10.248

Similarly as in the case of the *NED*1 measure, it has been found that the *XP* feature had the greatest impact on the EER value in signature recognition with the use of the Levenshtein method.

4. CONCLUSIONS

In this paper the method of feature selection approach which uses statistical significance testing to rank signature features based on their association with the result of signature recognition was used. The study focused on determination of a combination of dynamic features of signatures, the use of which in the normalized Levenshtein method allows obtaining the lowest error in signature recognition. The analysis proves that there is a statistical relationship between signature features and the error returned by the classifier based on the normalized Levenshtein method. From obtained results follow that the best features selection is given by combination of feature *X* and feature *P*. For these parameters the EER coefficient achieves the lowest values.

Next stages of the research will aim at comparing the result obtained by means of the test χ^2 with the results obtained with the use of other tests known from the literature. Also other features of signatures, different from those presented in this paper, will be taken into account.

BIBLIOGRAPHY

[1] CHA S., Comprehensive survey on distance/similarity measures between probability density functions, International Journal of Mathematical Models and Methods in Applied Sciences, 2007, Vol. 1(4), pp. 300-307.
[2] DOROZ R., WRÓBEL K., PORWIK P., Signatures recognition method by using the normalized Levenshtein distances, Journal of Medical Informatics and Technologies, 2009, Vol. 1, pp. 73-78.
[3] IMPEDOVO S., PIRLO G., Verification of handwritten signatures: an overview, 14th International Conference on Image Analysis and Processing (ICIAP'07), 2007, pp. 191-196.
[4] KHAN M.K., KHAN M.A., Khan M.A.U., Ahmad I., On-line signature verification by exploiting inter-feature dependencies, 18th International Conference on Pattern Recognition (ICPR'06, 2006), Vol. 2, pp. 796-799.
[5] LEVENSHTEIN V.I., Binary codes capable of correcting deletions, Insertions, And Reversals, Soviet Physics Dokl., 1966, pp. 707-710.
[6] MARZAL A., VIDAL E., Computation of normalized edit distance and applications, IEEE Trans. Pattern Analysis and Machine Intelligence, 1993, Vol. 15, No. 9, pp. 926-932.
[7] PARA T., MITAS M., Determining signatures' characteristic features using statistical methods, Journal of Medical Informatics and Technologies, 2008, Vol. 1, pp. 41-50.
[8] RABINER LAWRENCE R., A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings Of The IEE, 1989, Vol. 77, No. 2.
[9] SCHIMKE S., VIELHAUER C., DITTMANN J., Using adapted Levenshtein distance for on-line signature authentication, Proceedings of the 17th International Conference, 2004, Vol. 2, pp. 931-934.
[10] WEIGEL A., FEIN F., Normalizing the weighted edit distance, Proc. 12th IAPR Int'l Conf. Pattern Recognition, Conf. B: Computer Vision and Image Processing, 1994, Vol. 2, pp. 399-402.