Statistical Analysis in Signature Recognition System Based on Levenshtein Distance

Malgorzata Palys, Rafal Doroz, and Piotr Porwik

Abstract. In this paper we develop our previously presented studies, where adaptation of the Levenshtein method in a signature recognition process is proposed. 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. Results obrained were tested by means of χ^2 independence test. It allowed determining the relationship between signature features and the errors of classifier.

1 Introduction

In the modern world security problems are increasingly very important, because safety of goods, resources and data should be procected. In order to protect them, common methods based on human knowledge are used, 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. Signature verification from biometric features point of view presents some advantages, such as: non invasive, intuitive and fast, well accepted socially and legally. Additionally, signature verification generally has a low storage requirements. The effectiveness of the use of an analysis of handwritten signatures as a biometric technique is very high. The main

Malgorzata Palys · Rafal Doroz · Piotr Porwik
Institute of Computer Science, University of Silesia, ul. Bedzinska 39, 41-200 Sosnowiec
e-mail: {malgorzata.palys,rafal.doroz,piotr.porwik}@us.edu.pl

R. Burduk et al. (Eds.): *CORES 2013*, AISC 226, pp. 217–226.
DOI: 10.1007/978-3-319-00969-8_21

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 [2], [3], [6], [7]. It should be noted that verification and identification of objects is also proposed in other scientific problems [4], [8].

This study presents a method of comparing signatures with the use of the normalized Levenshtein metrics [9], [13]. The effectiveness of these metrics in the process of signature recognition has been examined. A large number of results was obtained, which made an analysis more difficult. Therefore, the presented method includes a detailed statistical analysis, which allows approach to selects features of the signatures being adequately compared.

2 Feature Preparation

Biometrics signature is one of the longest-known security techniques invented by humans. Signature has for many years adopted a form of determining the credibility (such as during operations related to running a bank account). Data collection process within a signature recognition process can be divided into two categories: static and dynamic. The static system collects data using *off-line* devices. A signature is put on paper and then is converted into a digital form with the use of a scanner or a digital camera. In this case, the shape of the signature is the only data source, without the possibility of using dynamic data. Signature recognition on the basis of photos does not protect against fraud. On the other hand, dynamic systems use *on-line* devices, which register, apart from the image of the signature, also dynamic data connected with it. The most popular *on-line* devices are graphics tablets. By using tablet, a signature can be recorded in the form of an n-point set. Fig 1. presents an example of signature S_i, captured by tablet.

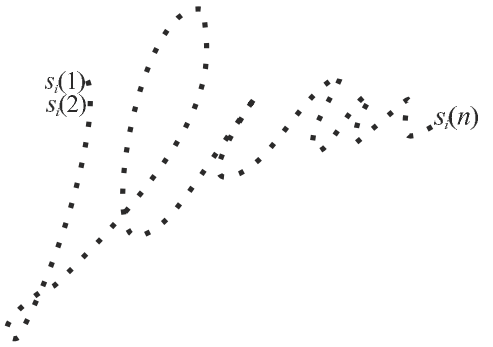

Fig. 1 An example of signature S_i

Signature S_i can be represented by the vector of points:

$$S_i = \begin{bmatrix} s_i(1) \\ s_i(2) \\ \vdots \\ s_i(n) \end{bmatrix}, \tag{1}$$

where:

$s_i(j)$ – the j-th point of signature S_i.

The tablet, during signing, is capable to measure many dynamics parameters, such as a pressure pen on tablet surface, position of pen, the angles at which the person holds a pen. This implies that there is a dynamic feature vector $c_i(u)$ associated with each point $s_i(u)$ of signature S_i:

$$s_i(u) \rightarrow c_i(u). \tag{2}$$

The local position (x_i, y_i) of the pen is given directly by a graphics tablet, while speed and acceleration can be obtained from this device or can be calculated on the basis of the position parameter. In presented work, following signature features were used:

$X = \{x_1, x_2, ..., x_n\}$, $Y = \{y_1, y_2, ..., y_n\}$ - sets of coordinates of n signature points,
$P = \{p_1, p_2, ..., p_n\}$ - set of pressure in particular points of signature,

Having the sets of mentioned features it is possible to determine, basing on them, additional features like as:

$V = \{v_1, v_2, ..., v_n\}$ - set of the speed of the pen in successive signature points [3],
$V_{up} = \{v_{up_1}, v_{up_2}, ..., v_{up_n}\}$ - set of the positive velocity values of the pen in successive signature points [5],
$V_{down} = \{v_{down_1}, v_{down_2}, ..., v_{down_n}\}$ - set of the negative velocity values of the pen in successive signature points [5],
$V_x = \{v_{x_1}, v_{x_2}, ..., v_{x_n}\}$ - set of the horizontal speed of the pen in successive signature points [3],
$V_y = \{v_{y_1}, v_{y_2}, ..., v_{y_n}\}$ - set of the vertical speed of the pen in successive signature points [3],
$P_{ch} = \{p_{ch_1}, p_{ch_2}, ..., p_{ch_n}\}$ - ses of the change in the pen pressure in successive signature points [5],
$K = \{k_1, k_2, ..., k_n\}$ - set of the inverse of the radius of the curve in successive signature points [12].

Because in proposed approach 10 signature features have been selected the extracted features form the vector:

$$c_i(u) = \left[c_i^1(u), c_i^2(u), ..., c_i^{10}(u)\right], \tag{3}$$

namely:

$$c_i(u) = [x_i(u), y_i(u), ..., k_i(u)]. \tag{4}$$

Finally, the signature can be described by the matrix S_i:

$$S_i = \begin{bmatrix} x_i(1) & y_i(1) & \cdots & k_i(1) \\ x_i(2) & y_i(2) & \cdots & k_i(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_i(n) & y_i(n) & \cdots & k_i(n) \end{bmatrix}. \tag{5}$$

3 Normalization of the Levenshtein Distance

The Levenshtein distance is the number of certain operations, called elementary operations, which must be performed to transform one character string into another one [2], [13].
Let $\sum$ define an alphabet of characters and a set containing all character sub-strings from this alphabet $\sum'$. Then, let's define two character strings and belonging to $\sum'$, where n and m are the lengths of these strings. Let $T_{A,B} = T_1, T_2, ..., T_l$ stand for the transformation of A character string into 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 $(a \to b)$. The sign λ represents 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 $(a \to b)$, $(b \to a)$.

Each elementary operation has a specific cost of its performance, which is called a weight of a given elementary operation. The weighting function δ assigns a non-negative real number to the i-th elementary operation $(a \to b)$:

$$\delta(T_i) = \delta(a \to b). \tag{6}$$

The weight of the $T_{A,B}$ transformation can be calculated using the following formula:

$$\delta(T_{A,B}) = \sum_{i=1}^{l} \delta(T_i). \tag{7}$$

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 $P_{A,B} = \left\{ P^1_{A,B}, P^2_{A,B}, ..., P^h_{A,B} \right\}$ 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:

$$W(P_{A,B}) = \delta(T_{A,B}). \tag{8}$$

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

$$GLD(A,B) = \min\{\delta(T_{A,B})\} = \min\{W(P_{A,B})\}. \tag{9}$$

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. $Ned1$ metric is defined by the formula:

$$Ned1(A,B) = \min\left\{\frac{W(P_{A,B})}{|P_{A,B}|}\right\}, \tag{10}$$

where:
$|P_{A,B}|$ – the number of elementary operations in an individual path. Another proposed measure is the $Ned2$ metric described by the following formula:

$$Ned2(A,B) = \min\left\{\frac{W(P_{A,B})}{|A|+|B|}\right\} = \frac{GLD(A,B)}{|A|+|B|}, \tag{11}$$

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

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

$$d_{N-GLD}(A,B) = \frac{2 \cdot GLD(A,B)}{\max(D,I) \cdot (|A|+|B|) + GLD(A,B)}, \tag{12}$$

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

All presented metrics: $Ned1$, $Ned2$, d_{N-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 coefficient, the metrics (13), (14) and (15) were adequately modified, so that the result of the comparison of two identical strings was the value 1:

$$NED1(A,B) = 1 - Ned1(A,B), \tag{13}$$

$$NED2(A,B) = 1 - Ned2(A,B), \tag{14}$$

$$NGLD(A,B) = 1 - d_{N-GLD}(A,B). \tag{15}$$

4 The Use of Normalized Levenshtein Metrics in the Process of Recognition of Handwritten Signatures

Levenshtein distance calculates the distances between two strings. In the presented method the values of individual feature of signatures S_i and S_j are compared. In order to use of normalized Levenshtein metrics in the process of recognition of handwritten signatures, new signature similarity measure LM_{ij}^{t} has been introduced. This measure is based on similarity metrics between two features t signature being compared. It will be described on the example of the $NED1$ metric:

$$LM_{ij}^{t} = NED1\left(S_i(t), S_j(t)\right), \tag{16}$$

where:
$S_i(t)$ –t-th column (feature) of signature matrix S_i.

At the beginning the values of signature features were normalized to the [0,1] range, so they can take indefinitely many values from this range [1], [10]. Therefore, the probability of occurrence of two identical feature values in two compared strings is near zero. In order to eliminate this situation, the additional parameter ϑ was introduced. It determines, to what maximum extent the two values being compared can differ from each other in order to be treated as equal. The features values $c_i^r(u) = c_j^r(u)$, if it fulfils the following condition:

$$\left|c_i^r(u) - c_j^r(u)\right| < \vartheta, \tag{17}$$

where:
ϑ – the maximum difference between the values of the features that allows recognizing them as equal,
$c_i^r(u)$ – the u-th element of r-th feature of the signature S_i,
$c_j^r(u)$ – the u-th element of r-th feature of the signature S_j.

The evaluation of the similarity of individual signatures was performed on the basis of an analysis of ten signature features and their combination. In order to specify the influence of a given feature on the result of the comparison, the weights w_1, w_2, w_3 of the features were introduced. Thus, 35 different values were obtained as the result of the comparison, and each of them described the similarity of a different combination of signature feature. The formula for determining the $WLM(S_i, S_j)$ similarity value of the two signatures S_i and S_j, taking into account combination of signature feature, is as follows:

$$WLM(S_i, S_j) = \left\{ LM_{ij}(w_1, w_2, w_3) : w_1, w_2, w_3 \in N \wedge \sum_{i=1}^{3} w_i = 1 \right\}, \tag{18}$$

where:

$$\forall_{\substack{k,l,m \in F \\ k \neq l \neq m}} \quad LM_{ij}(w_1, w_2, w_3) = LM_{ij}^{k} \cdot w_1 + LM_{ij}^{l} \cdot w_2 + LM_{ij}^{m} \cdot w_3 \tag{19}$$

N - weight of the signature feature, $N = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$,
F - number of feature, $F = \{1, 2, ..., 10\}$.
The individual element of the set F correspond to the number of columns of matrix S_i. For example, F = 1 means the feature X.

5 The Course and Results of the Studies

The studies were conducted for 50 signatures coming from different persons. The set of test signatures used in the studies comes from the SVC2004 database. The signatures were divided into 10 groups. Each group contained 4 original signatures of one person and 1 forged signature. In order to assess, which combination of signature feature has the greatest impact on EER values, the χ^2 test was applied. It allows determining whether there is a relationship between feature combinations and EER values.
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:

$$H_0 : P(Z = z_k \cdot U = u_m) = P(Z = z_k) \cdot P(U = u_m). \tag{20}$$

The alternative hypothesis H_1 shows a relationship between the Z and U:

$$H_1 : P(Z = z_k \cdot U = u_m) \neq P(Z = z_k) \cdot P(U = u_m), \tag{21}$$

where the variable Z is a combination of the signature features:

$$\begin{aligned} Z = \{&X, Y, P, V, V_{up}, V_{down}, V_x, V_y, P_{ch}, K, XY, XP, XV, XV_{up}, XV_{down}, \\ &XV_y, XV_x, XP_{ch}, XK, YP, YV, YV_{up}, YV_{down}, YV_y, YV_x, YP_{ch}, YK, \\ &XYP, XYV, XYV_{up}, XYV_{down}, XYV_y, XYV_x, XYP_{ch}, XYK\}. \end{aligned}$$

The variable U is a range of EER values. As the number of results for each of the three analysed measures was very high (1736733), the analysed data were divided into 7 subsets. Each subset was assigned to 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 $NED1$ measure (in which the lowest value of EER = 1.161%, and the highest value of EER = 54.918%), 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	[1.161-8.841)
Very good	[8.841-16.521)
Good	[16.521-24.201)
Average	[24.201-31.881)
Poor	[31.881-39.561)
Bad	[39.561-47.241)
Very bad	[47.241-54.921)

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. Then the expected quantities was calculated [11]. The seven feature combinations of the largest differences calculated between actual quantities and expected quantities were presented in Table 2 .

Table 2 Table showing the difference between the actual quantities and expected quantities of EER values for the *NED*1 measure

	XY	XV	XP_{ch}	XYP	XYV_{up}	XYV_y	XYV_x
Excellent	1366.90	-1033.10	-1038.83	3611.10	1265.38	1008.65	1931.93
Very good	64.54	-139.46	-501.12	80.22	342.56	-5.10	82.25
Good	-642.35	440.65	93.87	-1530.14	-619.92	-618.69	-770.47
Average	-488.51	534.49	596.59	-1227.37	-529.27	-338.17	-720.07
Poor	-228.56	170.44	634.48	-691.33	-331.29	-90.24	-413.20
Bad	-69.06	28.94	210.95	-232.12	-126.11	36.90	-107.08
Very bad	-2.95	-1.95	4.05	-10.36	-1.36	6.65	-3.35

For the *NED*1 measure, the calculated statistic is $\chi^2 = 32356.1$. The critical value χ^2_α =238.32 was taken from the distribution tables χ^2 for the adopted level of significance $\alpha = 0.05$. The quantity table has 7 rows and 35 columns, so $s = (7-1)(35-1)$ = 204 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 2 , it can be stated that the greatest impact on the EER value in the Levenshtein method has a combination of XYP 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. Table 3 showing the difference between the actual quantities and expected quantities of EER values for the *NED*2 measure whereas Table 4 for the *NGLD* measure.

Table 3 Table showing the difference between the actual quantities and expected quantities of EER values for the *NED*2 measure

	XY	XV_{down}	XV_y	XK	XYP	XYV_{up}	XYV_x
Excellent	1749.63	-1199.49	-1218.34	-1199.14	4622.21	1619.68	2472.87
Very good	86.48	168.22	-67.16	168.68	97.86	417.92	100.34
Good	-610.23	546.87	249.73	547.08	-1912.67	-774.90	-963.09
Average	-561.78	172.60	242.39	171.87	-1595.58	-688.05	-936.09
Poor	-205.71	53.44	442.82	53.52	-394.06	-188.83	-235.52
Bad	-110.50	-8.10	205.78	24.87	-336.58	-182.86	-155.27
Very bad	-2.39	-0.77	10.57	-0.80	-8.70	-1.14	-2.82

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

	XY	XV_{up}	XP_{ch}	YP_{ch}	XYP	XYV_{up}	XYV_x
Excellent	2152.04	-1333.68	-1595.64	-1335.05	5546.65	1943.62	2967.44
Very good	76.97	116.32	-765.51	-810.91	111.56	476.43	114.39
Good	-488.19	213.29	71.34	32.92	-1472.76	-960.87	-1194.23
Average	-449.43	202.96	425.61	796.12	-1818.96	-784.38	-1067.14
Poor	-242.73	275.70	1272.77	434.33	-464.99	-171.84	-214.33
Bad	-114.92	68.12	571.41	607.07	-400.53	-217.60	-184.77
Very bad	-2.49	5.09	3.41	2.38	-6.09	-1.31	-3.24

Statistics for the *NED*2 and *NGLD* measures are respectively $\chi^2 = 96432.7$ and $\chi^2 = 78784.6$. Thus, they belong to the same critical area as the *NED*1 measure. Similarly as the *NED*1 measure, it has been found that the *XYP* feature had the greatest impact on the EER value in signature recognition with the use of the Levenshtein method.

6 Conclusions

In this paper the method of feature selection with statistical significance testing was proposed. The study focused on determinating a combination of dynamic features of signatures which 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 XYP. For these parameters the EER coefficient achieves the lowest values. In the future the result obtained by means of the test χ^2 will be compared with other tests known from the literature. Also other features of signatures will be taken into account.

References

1. Cha, S.: Comprehensive survey on distance/similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences 1(4), 300–307 (2007)
2. Doroz, R., Porwik, P.: Handwritten signature recognition with adaptive selection of behavioral features. In: Chaki, N., Cortesi, A. (eds.) CISIM 2011. CCIS, vol. 245, pp. 128–136. Springer, Heidelberg (2011)
3. Doroz, R., Wróbel, K.: Method of signature recognition with the use of the mean differences. In: Proceedings of the ITI 2009 31st International Conference (ITI 2009), pp. 231–235 (2009)
4. Froelich, W., Wakulicz-Deja, A.: Probabilistic Similarity-Based Reduct. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 610–615. Springer, Heidelberg (2011)
5. Gupta, G.K.: The state of the art in on-line handwritten signature verification4 (2006)
6. Impedovo, S., Pirlo, G.: Verification of handwritten signatures: an overview. In: 14th International Conference on Image Analysis and Processing (ICIAP 2007), pp. 191–196 (2007)
7. Khan, M.K., Khan, M.A., Khan, M.A.U., Ahmad, I.: On-line signature verification by exploiting inter-feature dependencies. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 2, pp. 796–799 (2006)
8. Koprowski, R., Wrobel: The cell structures segmentation. In: 4th International Conference on Computer Recognition Systems (CORES 2005), pp. 569–576 (2005)
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, Insertions, And Reversals. In: Soviet Physics Dokl, pp. 707–710 (1966)
10. Marzal, A., Vidal, E.: Computation of normalized edit distance and applications. IEEE Trans. Pattern Analysis and Machine Intelligence 15(9), 926–932 (1993)
11. Para, T., Mitas, M.: Determining signatures characteristic features using statistical methods. Journal of Medical Informatics and Technologies 1, 41–50 (2008)
12. Pastor, M., Toselli, A., Vidal, E.: Writing speed normalization for on-line handwritten text recognition. In: Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, pp. 1131–1135 (2005)
13. Schimke, S., Vielhauer, C., Dittmann, J.: Using adapted Levenshtein distance for on-line signature authentication. In: Proceedings of the 17th International Conference, vol. 2, pp. 931–934 (2004)
14. Weigel, A., Fein, F.: Normalizing the weighted edit distance. In: Proc. 12th IAPR Intl Conf. Pattern Recognition, Conf. B: Computer Vision and Image Processing., vol. 2, pp. 399–402 (1994)