Analysis of Geometric Features of Handwriting to Discover a Forgery

Henryk Maciejewski and Roman Ptak

Wrocław University of Technology,
ul. Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland
e-mail: {Henryk.Maciejewski,Roman.Ptak}@pwr.wroc.pl

Abstract. This work proposes a method of analysis of geometric features generated from hand-written text to verify a supposition that a given sample of text of unclear authorship (e.g., a signature or initials) and some given reference text of known authorship have been written by the same author. The method is targeted to problems where the reference material is relatively large and the sample of unclear authorship is small, hence the number of feature vectors for the two groups compared is highly unbalanced. This makes the problem computationally challenging as standard approaches based on statistical hypothesis testing to compare distributions cannot be used. We propose a method to estimate the likelihood that the set of features observed in the small sample comes from the distribution generated from the reference material. This approach can be used to help discover or prove a forgery in documents.

1 Introduction

Analysis of handwritten text has been an important application area of machine learning or advanced statistical pattern recognition since early years of these disciplines. The purpose of analysis has been primarily the *recognition* of handwriting or personal *identification* of the writer based on hand-writing [4,6]. Various approaches to text recognition and writer identification were proposed in literature, e.g., based on Hidden Markov Models (HMM), Support Vector Machines with specialized kernels, or based on combining several different classifiers, [1,2,7,8]. Two different approaches to recognition were developed: *offline* or *online*, depending on whether features are derived only from the handwritten text (offline methods), or maybe also based on the analysis of the very process of writing (online methods).

The purpose of the method developed in this work is the detection of *forged hand-writing*, e.g., forged signatures or initials, etc. We assume that a signature (or a few signatures) is available, for which the authorship is unclear or questioned. We denote this hand-written text as the Questioned Material (QM). We also assume that a large reference sample is available for which the authorship is known; we denote this as the Reference Material (RM). QM and RM are schematically

W. Zamojski et al. (Eds.): Dependable Computer Systems, AISC 97, pp. 145–154.
springerlink.com

illustrated in Fig. 1 and 2. The proposed method aims to verify the hypothesis that the QM and RM were written by the same author. We develop an offline method based on the analysis of geometric features of handwriting. It should be noted that in this problem formulation, the number of features derived from the QM is inevitably small, while the number of features from the RM is large enough to allow for estimation of the distribution of features in RM. The method proposed is intended to compare features between such highly unbalanced classes.

In the following section, we define the set of geometric features that can be computed on the basis of short hand-written texts, such as signatures or initials. Next we propose a method to verify the hypothesis that feature vectors generated from QM and RM come from the same distribution. If the hypothesis proves true, we conclude that QM was not forged, i.e. it was written by the person who wrote RM, otherwise QM is interpreted as forged. We provide a numerical example to illustrated this approach.

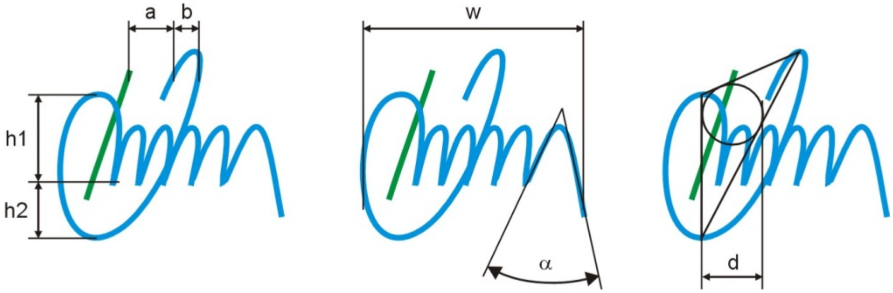


Fig. 1 The initials model of the questioned material (QM) and measured features (explanation in text)

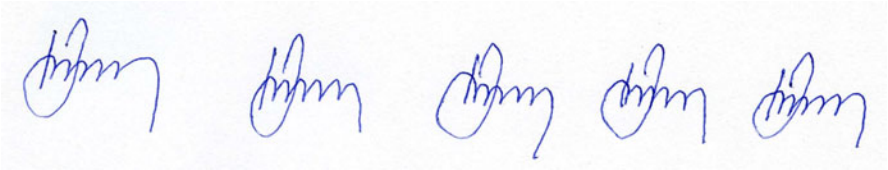


Fig. 2 A sample of text in the reference material (RM)

2 Geometric Features Derived from Handwritten Text

When performing a handwriting examination and comparison, many features of the handwriting are taken into consideration. It is possible to obtain features at the following document levels: basic, macrostructural and microstructural. The Catalogue of Graphic Features of Handwriting distinguishes five groups of features: synthetic, topographic, motor, measurable and constructional [3,5]. Some of the features of handwriting are: the overall size of the writing, the width of letters, words, etc., pen lifts within and between letters, the curvature of pen strokes [4]. Geometric features are parts of topographic features and measurable features on basic document level. Some of these features are used also for examination of long texts but are also applicable for short texts (such as initials), examples are shown in Figs. 3 and 4. Other features are specific to examination of initials (e.g., the radius of rotation of the hand, illustrated in Fig. 1 and explained in the following paragraph).

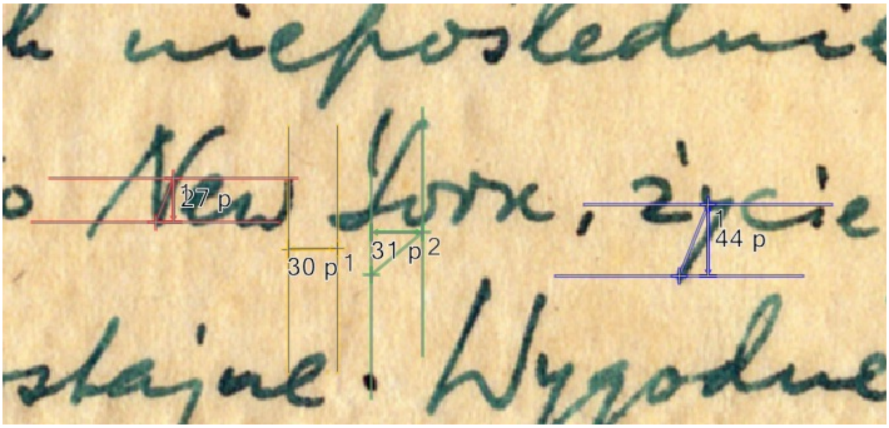


Fig. 3 Examples of features measured from long text: the size of handwriting (the high of middle zone), the high of lower zone, distance of words and distance between strokes in pixels

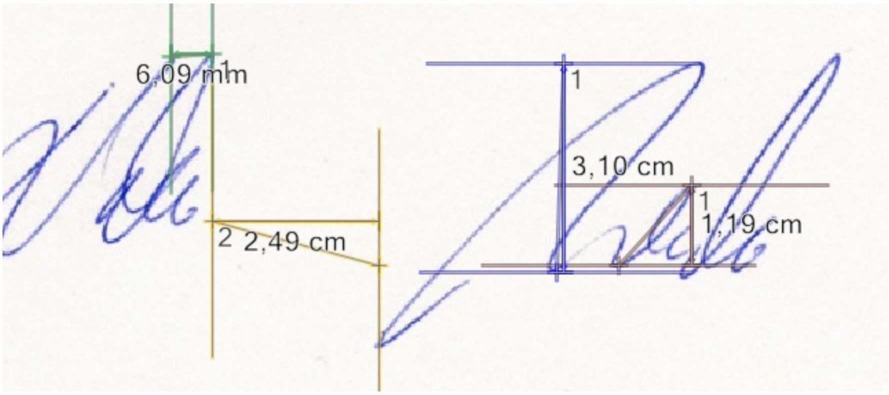


Fig. 4 Example of features measured from short text (two initials): the high of middle zone, the high of upper zone, distance of words and distance between strokes in mm and cm

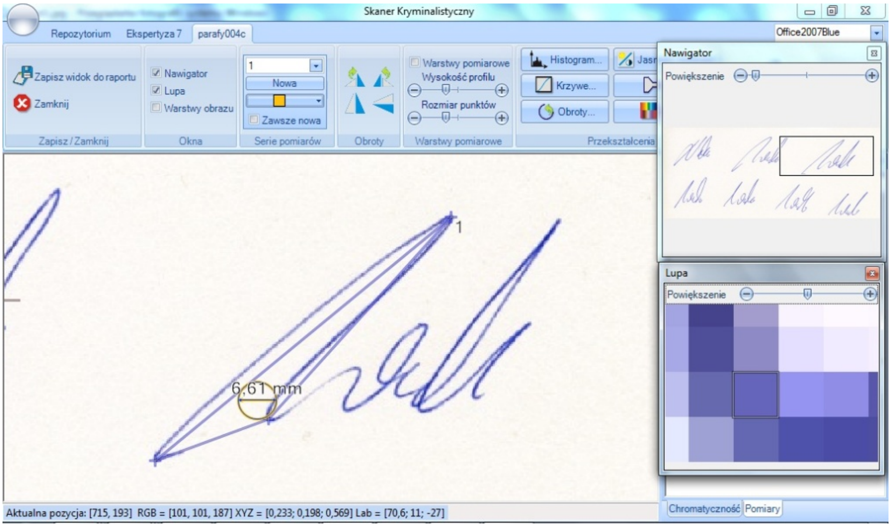
The proposed method of analysis of short text will be illustrated using the sample initials shown in Fig. 1. The initials consist of two strokes. The variable a represents between strokes distance and b represents inner main strokes distance. Feature h1 is the high of middle zone and h2 is the high of lower zone. The next three extreme points of the main stroke of initials mark the limits of pen trajectory. The inscribed circle in the triangle represents the radius of rotation of the hand.

This feature is also dependent on the human anatomy. We also measure the angle α between final strokes of initials. Proportions between the various features of handwriting (ratios) are generally considered most informative.

We propose to use the following geometric features to compare the questioned material with the reference material:

- Feature 1: f_1=a/b is the ratio of the outer (a) to inner (b) strokes distance,
- Feature 2: f_2=h1/h2 is the ratio of the high of middle zone (h1) to the high of the lower zone (h2),
- Feature 3: f_3=α is the angle of the final strokes,
- Feature 4: f_4=w/d is the ratio of the initials width (w) to the diameter (d) of an inscribed circle of a triangle determined by the extreme points in main stroke of initials.

In order to illustrate the proposed method we will analyze the set of 120 initials in the RM class and two initials in the QM class as schematically shown in Figs. 1 and 2. The geometric features were measured for these texts in a semiautomatic way using the developed software application shown in Fig. 5.



Aktualna pozycja: [715, 193] RGB = [101, 101, 187] XYZ = [0,233; 0,198; 0,569] Lab = [70,6; 11; -27]

Fig. 5 Measurement window of the software application developed to facilitate examination of documents

Values of the proposed geometric features calculated from the two initials in the QM group are summarized in Table 1.

Table 1 Geometric features calculated from the two signatures in QM

Initials	A	b	**a/b**	h1	h2	**h1/h2**	**α**	w	D	**w/d**
QM1	180	260	**0.69**	503	517	**0.972**	**52.1**	1470	574	**2.56**
QM2	198	266	**0.74**	594	278	**2.136**	**49.1**	1386	584	**2.37**

In Fig. 6 we provide a preliminary comparison of the individual features between the QM and RM groups. The distribution of a feature in RM is represented by a boxplot, with the box representing the inter-quartile range (IQR) and the whiskers spanning the range ±1.5 IRQ around the median. It can be observed that QM and RM are most differential in terms of the f_3 feature which is significantly larger in QM than in RM (more specifically, the value of f_3 for the two samples in QM lie in the outlier area of the distribution of RM).

In the next section we propose a measure which aggregates the similarities between QM and RM in terms of individual features into a value of likelihood of randomly selecting a given QM sample from the distribution of RM samples.

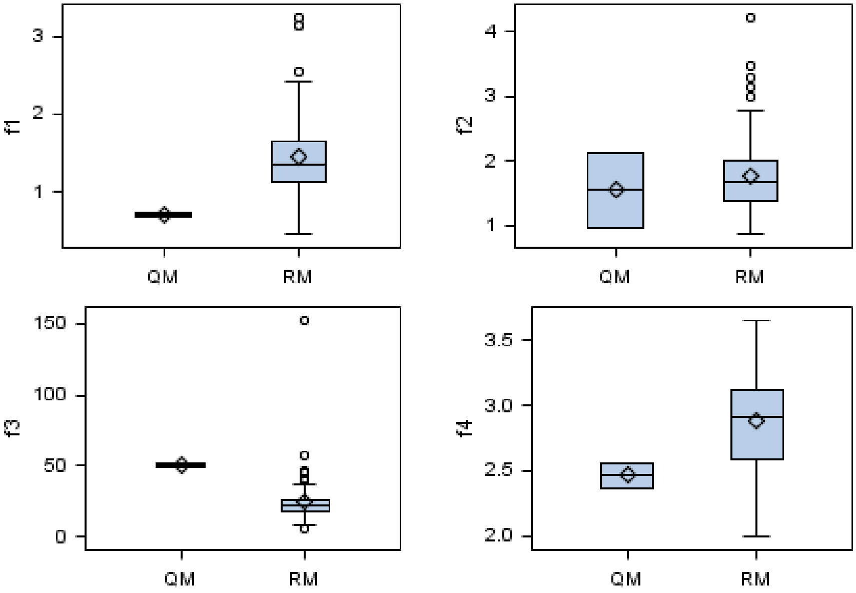


Fig. 6 Values of features from two samples of the questioned material (QM) shown in Table 1 compared with distributions of features calculated from the reference material (RM)

3 Verification of the Hypothesis of Common Authorship of QM and RM

In this section we propose a method to estimate the probability that a given sample (initials) from the QM group can be observed in the RM group. This measure is based on the features f_1 to f_4 of the QM sample compared with the corresponding distributions in the RM class.

We propose the following procedure.

1. For each of the features f_i, i=1,…,4 of the given sample from QM, we estimate the probability (denoted pc_i) that *the value f_i or a more extreme value* can be observed in the distribution for the RM class:

$$pc_i = \begin{cases} q_i & dla\ q_i < 0.5 \\ 1 - q_i & dla\ q_i \geq 0.5 \end{cases}$$

where $q_i = F_i(f_i)$, and F_i denotes the cumulative distribution function of the feature i estimated for RM. The value of pc calculated for q<0.5 is illustrated in Fig. 7 (pc_1=0.017, calculated for initials QM1, this value is represented by left tail shaded in red). The value of pc corresponding to q>0.5 is illustrated by the right tail in Fig. 8 (pc_3=0.12 calculated for initials QM1).

2. We estimate the joint probability that the set of features f_1-f_4 for the QM sample can be observed in the RM class:

$$p = \prod_{i=1}^{4} pc_i$$

This formula resembles the rule used in naïve Bayes classifier, where the features are assumed as independent. In this example, we observe that the features f_1-f_4 are weakly correlated (with the correlation coefficient ranging from 0.026 to 0.16) and realize very similar eigenvalues of the feature correlation matrix (results not shown) – this indicates weak association between features.

3. We calculate the number of samples (initials) in the RM class for which the value p (computed according to the formula in step 2) does not exceed the value p computed for the given QM sample. This number divided by the total number of samples in RM can be interpreted as the likelihood that the sample QM can be observed in the distribution of RM, under the hypothesis that QM comes from the distribution RM. We denote this as pVal:

$$pVal = \frac{1}{|RM|} \sum_{r \in RM} I(p_r \leq p)$$

where p_r is calculated for a sample $r \in RM$ according to step 2, and p is calculated for the sample QM according to step 2. The function I returns 1 if the condition is true, and 0 – otherwise.

To illustrate the way to verify the hypothesis that the QM initials have been written by the author of RM, we summarize the values q_i, p and pVal calculated by the proposed procedure in Table 2. With the confidence level of 5% we conclude that the sample QM2 was *not* written by the author of RM (as pVal<0.05 for QM2,

Table 2 The value of q for the initials in the QM class, and the probability of QM initials occurring in the RM class

Initials	q_1	q_2	q_3	q_4	p	**pVal**
QM1	0.017	0.287	0.88	0.18	0.00011	**0.0952**
QM2	0.029	0.822	0.87	0.07	4.82E-5	**0.0397**

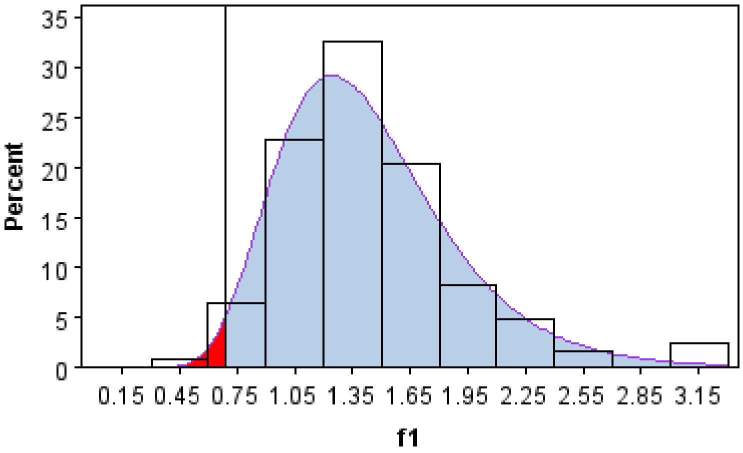


Fig. 7 Illustration of the value pc_1 for feature f_1 calculated for the initials QM1 from the distribution of this feature for RM. The pc value corresponds to the left tail of the distribution (marked red)

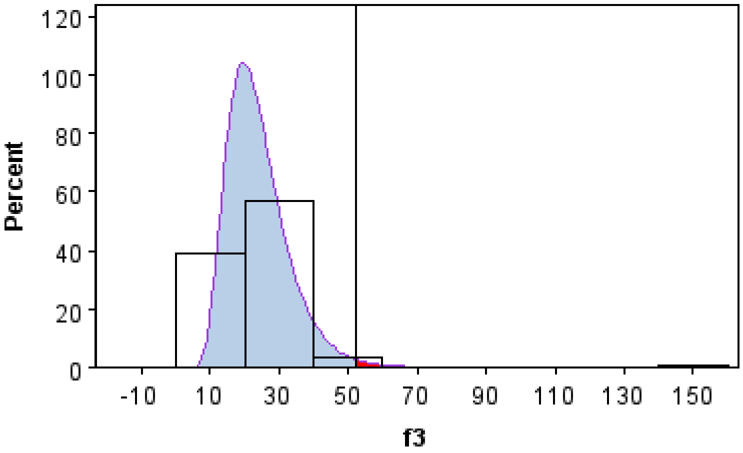


Fig. 8 Illustration of the value pc_3 for feature f_3 calculated for the initials QM1 from the distribution of this feature for RM (the pc values)

hence the hypothesis of common authorship is rejected). However, this conclusion cannot be formulated for the sample QM1 (as pVal > 0.05).

The pVal indicates how unlikely the value of p calculated for the QM sample is in the distribution pertaining to the RM class. This can be also shown graphically – see Fig. 9, where the p for QM (given in log scale) is compared with the distribution for RM.

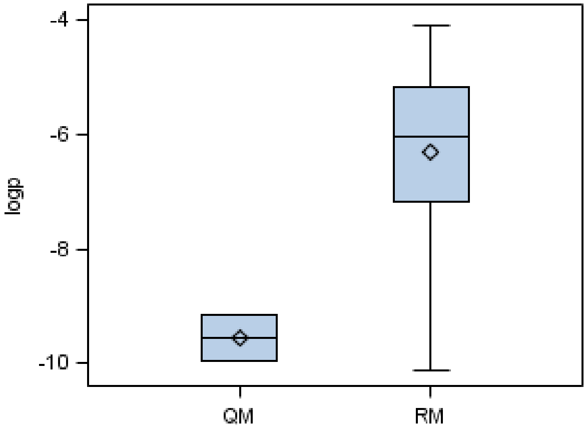


Fig. 9 The measure p calculated for the QM samples compared with the distribution for the RM group

The analyses discussed in this section and summarized in Table 2, Figs. 9 and 6 provide numerical and graphical indications about how *different* the questioned material seems to be as compared with the reference material. However, this approach is unable to discover text forgeries which consist in taking an (almost exact) copy of reference initials (using e.g., photocopying techniques). The method developed in the next section aims to discover this type of forgery by revealing that the questioned material should be *too similar* to the reference material.

4 Analysis of Similarity of Feature Vectors

In this section we extend the previous analysis by directly comparing feature vectors between the classes QM and RM. The purpose of this is to measure the *natural variability* of features in the RM group, characteristic of the set of initials written by a human (and not a machine). Then, if some of the QM initials are *too similar* to some of the RM samples, this clearly indicates possibly forged handwriting.

This idea motivates the following procedure.

1. For each of the samples $r_i \in RM$ from the reference group RM, calculate the distance $dmin_i$ to the nearest sample from RM:

   $$dmin_i = \min_{j \neq i}\{dist(r_i, r_j): r_j \in RM\}$$

   In the following example we use the Euclidean distance $dist(r_i,r_j)$ between the feature vectors.
2. For the QM sample $qm \in QM$ calculate the minimum distance $dmin_{qm}$ to the samples from RM, i.e.

$$dmin_{qm} = \min_{j}\{dist(qm_i, r_j): r_j \in RM\}$$

3. Estimate the probability pr= G(dmin_{qm}) that the value dmin_{qm} *or smaller* is observed in the distribution of dmin calculated from RM samples. If this probability is small, the QM sample seems *too similar* to some of the RM samples then expected taking into account the natural variability among RM samples. Hence the QM sample is presumably a copy of a RM sample (hence forged?).

   If pr is *not* small, then calculate the probability pr2=1- G(dmin_{qm}) that the value dmin_{qm} *or bigger* is observed in the distribution of dmin calculated from RM samples. If this probability is small, the QM sample seems *too different* from the RM samples then expected under the hypothesis that QM and RM samples come from the same distribution. Hence the QM sample is presumably written by a different author than RM.

To illustrate this, in Fig. 10, the distribution of dmin for the RM class is shown as a boxplot (right part of Fig. 10). We conclude that if a QM sample was a copy of a RM sample (i.e. dmin≈0), this would be immediately clear as the value of dmin≈0 is below the lower tail of the distribution.

The left part of Fig. 10 shows the values of dmin calculated for the two samples QM1 and QM2 analyzed in section 3. We clearly see that the value pr2 (see step 3) for these samples equals about 0 (as the values are far above the upper tail of the distribution). Thus we conclude that these two questioned samples were presumably written by a different author than RM.

This result is generally consistent with results obtained in section 3 (although here the QM1 sample is classified as different than RM, while the test in section 3 would need significance level of 0.1 to confirm this).

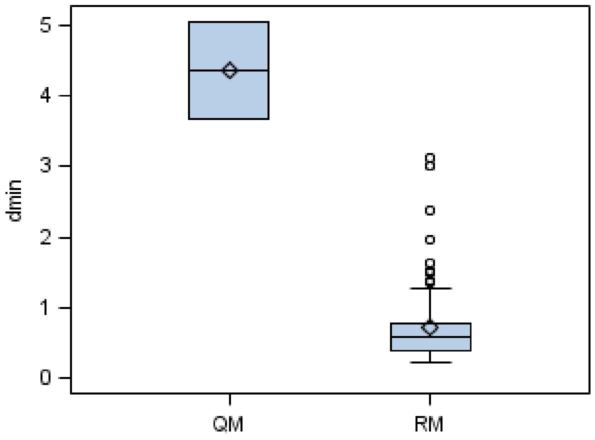


Fig. 10 Distribution of the distance to the nearest neighbor sample in the RM group (dmin) shown for the RM group (right boxplot), and distance of QM samples to the nearest RM sample (left boxplot)

5 Conclusions

The methods proposed in this work can be used to provide quantitative and graphical indications whether questioned text (such as initials) was written by the author of reference material. The analysis is based on the set of four geometric features calculated from samples of handwriting. The indications are based on probabilistic analysis of variability of features calculated from the reference material. The questioned initials are considered significantly different than the reference initials (which may indicate different authorship) if their features occupy the far right tail of the distribution characteristic of the reference group. We also propose a method to discover initials which are *too similar* to the reference material than expected taking into consideration natural variability in hand-written text. Such too similar initials may be deemed suspicious or forged.

However, it should be made clear that all results shown in this work are of probabilistic nature. As such, they should not be directly translated into a firm statement about some questioned material being forged. Such conclusions are to be made only by a trained human evaluator (a forensic expert), for whom the methods elaborated here may provide some decision support data.

References

[1] Bahlmann, C., Haasdonk, B., Burkhardt, H.: Online handwriting recognition with support vector machines - a kernel approach. In: Proc of the Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 49–54 (2002), doi:10.1109/IWFHR.2002.1030883

[2] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction. Springer, Heidelberg (2001)

[3] Katalog Graficznych Cech Pisma Ręcznego (The Catalogue of Graphic Features of Handwriting) (in Polish) (2007), `http://prawo.amu.edu.pl/uploads/slownik/aneks.htm` (accessed March 12, 2011)

[4] Koziczak, A.: Metody pomiarowe w badaniach pismoznawczych (Measurement methods in examination of handwriting, in Polish), Instytut Ekspertyz Sądowych, Kraków (1997)

[5] Morris, R.: Forensic handwriting identification. Fundamental concepts and principles. Academic Press, New York (2000)

[6] Saferstein, R.: Criminalistic: An Introduction to Forensic Science, 8th edn. Prentice-Hall, Englewood Cliffs (2004)

[7] Schlapbach, A., Bunke, H.: A writer identification and verification system using HMM based recognizers. Pattern Analysis & Applications 10(1), 33–43 (2007), doi:10.1007/s10044-006-0047-5

[8] Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans on Systems, Man and Cybernetics 22(3), 418–435 (2002)