Brazilian Forensic Letter Database

Cinthia O. A. Freitas[1] *Luiz S. Oliveira*[1] *Robert Sabourin*[2] *Flávio Bortolozzi*[3]
[1]Pontifical Catholic University of Paraná (PUCPR) – Curitiba – Paraná – Brazil
[2]École de Technologie Supérieure (ETS) – Montreal – Canada
[3]OPET Faculty (OPET) – Curitiba – Paraná – Brazil
{cintha, soares}@ppgia.pucpr.br, robert.sabourin@etsmtl.ca, flavio.bortolozzi@opet.com.br

Abstract

In order to make writer identification a more consistent area of research, several authors have made efforts to build complete and concise forms in the sense that all letters and numerals would be captured. However, most of the works found in the literature are devoted to the English language, thus, they do not contemplate certain particularities of the Brazilian Portuguese language. This paper describes the Brazilian Forensic Letter database considering the image properties and important graphometric features such as, writing style, relative slant, relative relationship between letters and baseline, and relative placement habits. The database has 315 writers and we are continuing to collect data.

Keywords: Forensic Letter, Writer Identification.

1. Introduction

Forensic document examiners (FDE), criminologists, and investigators have long been challenged to use handwriting inscriptions or markings as clues to identify their writers. The main purpose is to find evidences that testify the originality of the document and association/dissociation among the authors and the handwriting documents. This kind of problem is relevant to understand handwriting and questioned documents in several cases, including credit card voucher fraud or bank checks fraud.

Writing is considered a brain function. The hand, foot or, mouth are merely devices with which to carry out instructions sent to it by the brain [9]. So, the human being develops writing skills since the childhood, going through the adolescence until reach the movements associated with "graphic maturity" or writer's "master pattern" [9].

When the goal is the forensic handwriting identification we need to take into consideration the following major principles of handwriting identification, summarized as follows [9]:

- No two people write exactly alike;
- Nobody writes exactly the same way twice;
- The significance of any feature, as evidence of identity or non-identity, and the problem of comparison becomes one of considering its rarity, the relative speed and naturalness with which it is written, and its agreement or disagreement with comparable features;
- No one is able to imitate all features of another person and simultaneously write the same relative speed and skill level as the writer he/she is seeking to imitate;
- For those writings where the writer successfully disguises his/her normal handwriting habits or where he/she imitates – traces – the writing habits of another writer while leaving no trace of his own, it is virtually impossible to identify the imitator.

We need to remember that there are different factors influencing letter formation, and that it is a result of the method used to learn to write. After practicing letter formation by using different strokes, the writer's writing speed increases and take particularities from the writer individuality, as shown in Figure 1. The individuality of handwriting has been studied and important works have been developed by Srihari [14,15,16].

Figure 1. Writer individuality: name "Fernando" written by two different writers.

To make writer identification a more consistent area of research, several authors have made efforts to build complete and concise forms in the sense that all letters and numerals would be captured [11,14]. However, most of the works found in the literature are devoted to the English language, thus, they do not contemplate certain particularities of the Brazilian Portuguese language.

In light of this, our objective in this paper is to present the Brazilian forensic letter database [6], which was created to address the several particularities of the Portuguese language, such as diacritics (á, à, ã, ê, ü) and the special symbol (ç), as presented in Figure 2. The database is composed of handwriting samples of 315 different writers. In addition, we present some graphometric features extracted from this database to give a better insight in the data.

The paper is divided into five sections. Section 2 introduces the topic forensic letters and databases. Section 3 presents an overview about the Brazilian

forensic letter database, which we call the PUCPR Letter. Section 4 summarizes some experimental results obtained with Brazilian Forensic Letter. Finally, in the Section 5 the discussions and future works are presented.

Figure 2. Portuguese language particularities.

2. Forensic Letters

There are different forms for forensic letters, which are applied to different countries. All of them, though, are in English, which makes it not complete suitable for studies involving the Brazilian Portuguese language. For example, the letters like “k”, “w”, and “y” are used in Brazil only for personal names.

The main goal of a forensic letter is to reproduce the association among different letters, words, numerals, and symbols. The letters must be adapted to the local language and as stated before, it is important to reproduce the upper and lower case, diacritics (á, à, ã, ê, ü), special symbols (ç), numerals (“0” to “9”), and punctuation symbols.

From the literature, there are several forensic letters devoted to collect the handwriting style, individuality, and other characteristics (static and dynamic) [7,11]. The forensic letters have been collected for several reasons. In their practice, document examiners frequently have to collect specimen of handwriting to make comparisons. The presentation of scientific testimony in general and handwritten document examination testimony in particular has aroused much interest. To ensure that in the sample obtained the writer writes all the letters of the alphabet (both in lower and upper case), as well as numerals, the so called "London Letter" was devised by Osborn [11]. We all know about the famous pangram "*A quick brown fox jumps over the lazy dog*", but the London letter is a kind of "*Superpangram*" giving all letters both in lower and upper case as well as all nine numerals [9].

Some examiners however feel that the language in the "London Letter" is too artificial and stilted, posing difficulties for the writer and requiring excessive concentration on the text (particularly when dictated but permissible to present in printed or typewritten form), so over the years several variations have been developed. An example of an "improvement" is the "Idaho Letter" that was published by Osborn in the 1920’s [11].

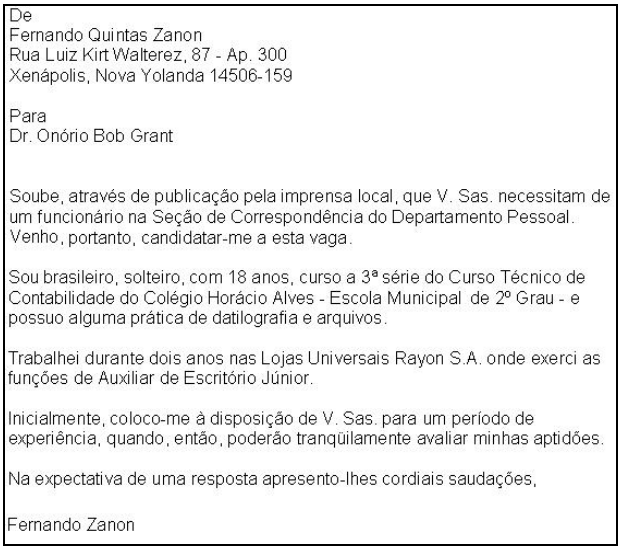
It is important remark that there are many different types of forms in different languages, and no one form is better or worst than any other specimen writing forms. Each has its specific purpose and is generally designed by an Forensic Document Examiners who believes that it will provide him with those features of the writer’s writing that are most useful in conducting a handwriting examination [9].

3. Brazilian Forensic Letter Database

FDE are always concerned about two aspects during the comparisons. These aspects are the variability of handwriting within individuals and between individuals. The extraction of significant features from the individual writing is the goal on determining the authorship. If we build a database that captures these characteristics and extract the distinctive individual characteristics, we would be able to obtain conclusions about the writer’s individuality and, later, develop computer-based analysis of handwriting documents.

The PUCPR forensic letter is composed of all Latin alphabetic characters (upper and lower case) and others Latin symbols such as, special symbol (ç), punctuation symbols, and diacritics (á, à, ã, ê, ü) [6]. The special symbol (ç) and diacritics constitute the feature called minimal graphics categorized as genetic features in graphometrical features [10].

There are no restriction concerning to the handwriting style. The other forensic letter models were studied to create the text in the PUCPR letter, as shown in Figure 03 [6].



De
Fernando Quintas Zanon
Rua Luiz Kirt Walterez, 87 - Ap. 300
Xenápolis, Nova Yolanda 14506-159

Para
Dr. Onório Bob Grant

Soube, através de publicação pela imprensa local, que V. Sas. necessitam de um funcionário na Seção de Correspondência do Departamento Pessoal. Venho, portanto, candidatar-me a esta vaga.

Sou brasileiro, solteiro, com 18 anos, curso a 3ª série do Curso Técnico de Contabilidade do Colégio Horácio Alves - Escola Municipal de 2º Grau - e possuo alguma prática de datilografia e arquivos.

Trabalhei durante dois anos nas Lojas Universais Rayon S.A. onde exerci as funções de Auxiliar de Escritório Júnior.

Inicialmente, coloco-me à disposição de V. Sas. para um período de experiência, quando, então, poderão tranqüilamente avaliar minhas aptidões.

Na expectativa de uma resposta apresento-lhes cordiais saudações,

Fernando Zanon

Figure 3. PUCPR forensic letter.

The PUCPR database has 315 writers, three letters per writer, in plain A4 sheets and there is no pen-drawn baseline. Table 1 presents the handwriting styles found in the database according to Tappert’s classification [17]. The letters were scanned in 300dpi, 256 grey levels. A binary version of the database is also available, as presented in Figure 04.

The PUCPR letter is concise (131 words) and complete in the sense that it contains all characters (letters and numerals) and certain character combinations of interest [6]. The number of occurrences in each position of interest in the text is shown in Table 2. In addition, the letter also contained punctuation (“.”, “,”), ten classes of numerals, special symbol (ç), and different combinations such as “nh”, “lh”, “qu”, and “00”, as reported in Table 3. These combinations are

very important grammatical elements that allow handwriting individuality studies, because few forgers take trouble to carefully simulate the lowercase letters. Normally, the forgers expend much energy in duplicating the capital letters believing that if they reproduce those correctly, the forgery will be accepted as a genuine writing [7].

As mentioned before, the significance of any feature is related to its rarity, relative speed and naturalness with which it is written, and its agreement or disagreement with comparable features. Moreover, the different combinations of characters ("nh", "lh", "qu" "00"), special symbols (ç), and diacritics (á, à, ã, ê, ü) can contribute for writer identification because the existence or not connecting strokes and how they are made (usage, curvature, direction of pen movement, etc) are the union of rarity, relative speed and naturalness [9]. The main question is: How many times I saw this pattern?

Table 1. Handwriting style.

Style	Distribution (%)
Boxed discrete character	7.9
Spaced discrete character	14.9
Run-on discretely written character	16.2
Pure cursive script writing	50.8
Mixed	10,2

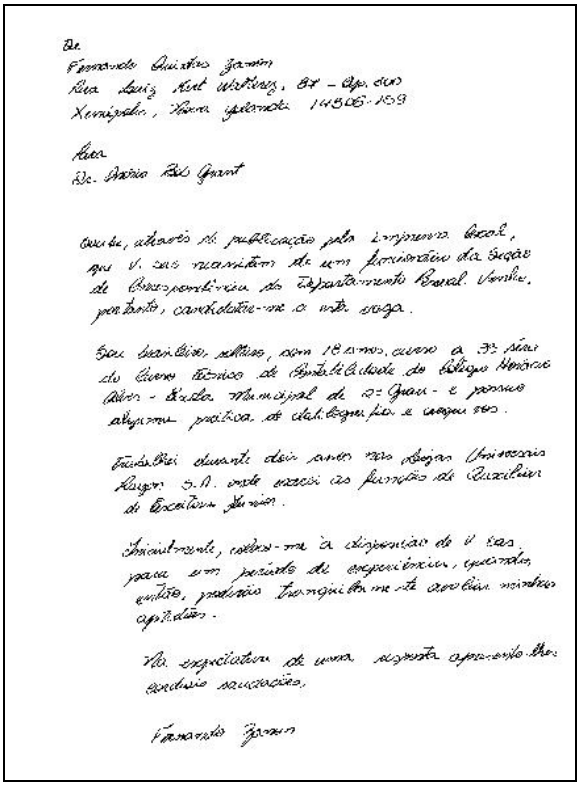


Figure 4. PUCPR letter - CF00078_01.

Research has shown that individuality and the act of writing are inseparable [9,10,11]. In handwriting identification, the task of FDE is to determine which features, when taken collectively, comprise the individuality of the writer [9]. Therefore, three concepts are important when the principal goal is writer identification: legibility, individuality, and identifiable writing. Each concept represents a particular feature of writing and a good example of how these concepts are applied to handwriting is found in Figure 5.

Table 2. Positional frequency of occurrence of letters.

upper	initial	Lower	start	mid	end
A	3	**a**	12	55	20
B	1	**b**	1	4	1
C	4	**c**	5	17	---[1]
D	2	**d**	18	15	---
E	2	**e**	7	41	21
F	2	**f**	2	1	---
G	2	**g**	0[2]	4	---
H	1	**h**	0	4	---
I	1	**i**	1	48	2
J	1	**j**	0	1	---
K	1	**k**	---	---	---
L	2	**l**	2	19	3
M	1	**m**	3	6	4
N	2	**n**	3	39	3
O	1	**o**	1	44	26
P	2	**p**	8	11	1
Q	1	**q**	1	2	1
R	2	**r**	1	38	5
S	6	**s**	3	20	20
T	2	**t**	1	24	2
U	1	**u**	3	18	2
V	4	**v**	1	7	---
W	1	**w**	---	---	---
X	1	**x**	0	4	0
Y	1	**y**	---	---	---
Z	2	**Z**	0	0	2

Observing the two capital characters in Figure 5a, "F" and "R", one could observe that both of them begin with a down stroke, a clockwise spiral movement up and to the right forming the right side of the letter. In the character "R" other spiral movement down is performed. Both letters are disconnected from the lower case letters "e" and "u". These are good examples that the beginning strokes and relative spacing are features related to legibility, individuality, and identifiable writing concepts when we compared the letter "F" and "R" between writers from Figure 5a and 5b.

[1] "---" denotes that this letter can not be found in this specific position in Brazilian Portuguese language.

[2] "0" denotes that this letter can be found in this specific position in Brazilian Portuguese language but there is no words in Brazilian forensic letter including this letter in this specific position.

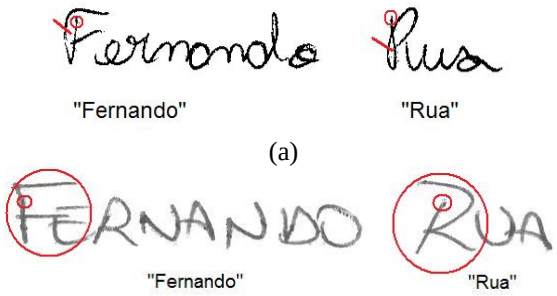


Figure 5. Words written by two diferent writers: a) letter CF00004_03 and b) CF00002_01.

Table 3. Positional frequency of occurrence of numbers, diacritics, puntuactions, and combinations.

number		**diacritic**	
1	3	**á**	4
2	1	**à**	1
3	2	**ã**	5
4	1	**é**	4
5	2	**ê**	2
6	1	**í**	1
7	1	**ó**	2
8	2	**õ**	3
9	1	**ú**	1
0	1	**ü**	1
00	1	**ç**	5
punctuation		**combination**	
.	13	nh	2
,	14	lh	2
-	7	br	1
		gr	3
		pr	3
		tr	3
		qu	4

FDE are always involved about three topics. Firstly the requirements for collected specimens, secondly the comparisons, and finally but not less important knowing what is significant in the handwriting to be used in the writer identification task.

The requirements for collected specimens are faced by the common sources of collected specimen handwriting such as, banks (deposit form, signature card, loan application, cancelled cheque), personal documents (ID, passport, greeting card, diary), business and school (agenda, application form, notebooks).

About the comparison, the Forensic Document Examiners are concerned about the features of the handwriting and the establishment the significance of the feature. In handwriting comparison, variation has been mostly a qualitative rather than quantitative assessment [9]. The writer's ability to write each feature rapidly, coupled with the frequency of its occurrence in random writing is what gives each feature its significance.

From the expert's point of view the features are classified as general and specific features. The general features are categorical and describe qualitative characteristics as: degree of connection between letters, slant or slope, motion, elaboration or presence of ornaments, direction of movement (clock-wise or counter-clock-wise), average number of strokes used to draw separate letters [5].

Specific features are known as graphometric and are aimed at the automatic or semi-automatic measurement of the following characteristics: distances between rows, height and width of letters, distances between letters, size of the above-row and under-row elements, distances between words, predominant slant, geometric parameters of handwriting elements like strokes, fragments and/or combination of characters [5,10].

Based on these considerations, we analyzed different characteristics such as, ratio-relative relationships [9] and graphometric [10] such as relative slant, relative relationship between letters and baseline and relative placement habits. Ratio is the relationship between two or more objects, in this case two or more components of handwriting. These are global characteristics and are very important for writer identification. Moreover, these features allow us to obtain a better insight about the database a meaningful characterization of the Brazilian writing style. Our analysis were performed using visual inspection and a process of automatic measurements (relative placement habits).

3.1. Relative Slant

Slant is a graphometric feature well known in graphometry, which has been extensively applied to writer identification [2,9]. This feature is quite usefully for global analysis, because the overall or average slant of the writing usually is uniform for uppercase letters, lowercase letters, and the components of individual letters. The degree of overall slant is dependent upon the preferences of the writer, the copybook style of writing learned by the writer, the naturalness of his writing, and it is influenced by several factors such as, the position of the writer's arm, the way he holds his pen, and the angle of the paper [4,9,13].

In general, there are three broad categories: right (forward), vertical (straight up and down), and left (backward). Table 4 presents the distribution in the database considering the relative slant.

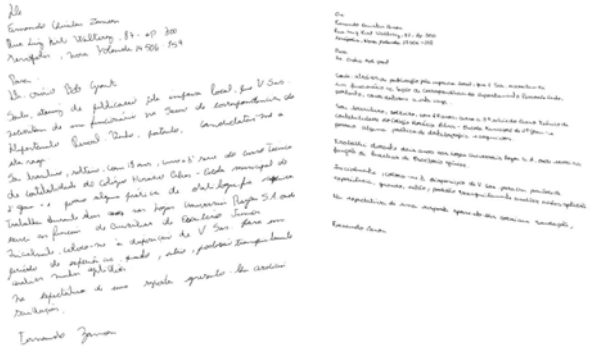
3.2. Relative Relationship Between Letters and Baseline

When one learns how to write, one of the lessons reinforced in the process is to write letters and words using the pen-drawn baseline [9]. Even when the baseline is imaginary, writers tend to adhere to this rule. When there is pre-printed baselines the writers can maintain the horizontal direction or rising or falling the

lines, as depicted in Figure 6. Table 5 presents the distribution in the database considering this individual characteristic.

Table 4. Relative slant.

Slant	Distribution (%)
Right (↗)	14.0
Vertical (↑)	80.3
Left (↖)	5.7



(a) **(b)**

Figure 6. Relationship between letters and baseline: a) rising (writer CF00025_01) and b) falling (writer CF00005_01).

Table 5. Relative relationship between letters and baseline.

Baseline	Distribution (%)
Rising	26.3
Horizontal	51.1
Falling	22.6

3.3. Relative Placement Habits

Placement of material on a form or a sheet of paper is significant for writing identification [7,9]. It can be observed in the PUCPR letter database that different writers start and stop their writing at different locations. Then, locations, such as sentence indentation, treatment and shape of margins, use of space, starting and stopping points; are examples of the relative placement habits [9].

We observed in the PUCPR letter database three different sentence indentation such as, left-right (Figure 4), top-down, and right-left, as presented in Table 6.

Table 6. Treatment and shape of margins.

Baseline	Distribution (%)
Left-right (\)	19.7
Top-down (|)	75.2
Right-left (/)	1.6
Mixed (>)	3.5

Other interesting characteristic is the use of space [10] by the writer. It must be remembered that any time the writer sees boundaries, his/her writing is probably going to be affected. In the PUCPR forensic letter there is no boundaries. But we observed that the writer delimited the space using imaginary boundaries as shown in Figure 7.

We computed the spacing not used by the writers and Table 7 presents the obtained results. We can conclude that in average almost 30% of the is not used by the writers. In other words, the writers use 70% of the space in the sheet of the paper (A4) to copy the PUCPR letter. The maximum value (72.1%) observed in region "D" represents that 70% of the total area not used to copy the letter. This value belongs to CF00255 writer. Analyzing the letters copied by this writer we observed that he has a very small relative height relationship between letters (upper and lower case) and a very small width relationship between letters (from the starting point of the approach stroke to the foot of the right side of the letter where it makes contact with the imaginary baseline).

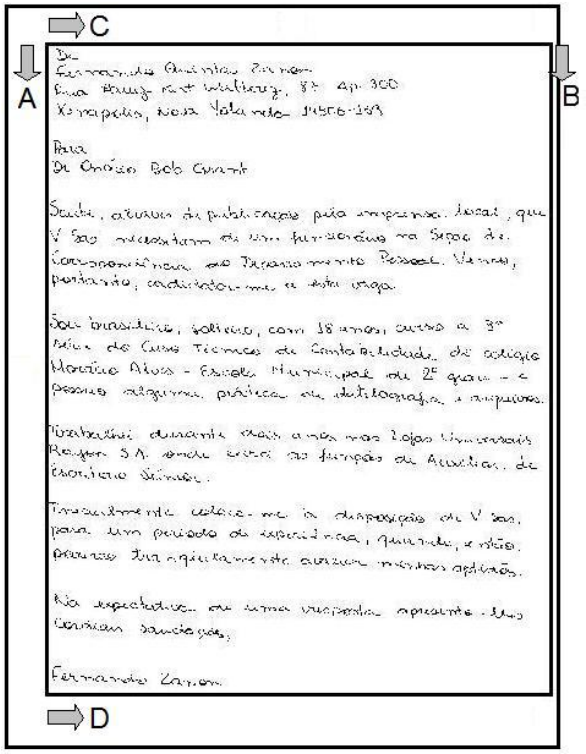


Figure 7. Use of spacing in the sheet of paper (A4 size, writer CF00082_02).

Table 7. Use of space in the sheet of paper (%).

Statistic values	A	B	C	D	Total
Average	3,0	6,0	5,3	18,6	32,9
Standard deviation	3,5	3,6	3,1	10,2	12,8
Mode	0,4	2,2	2,8	16,2	22,3
Maximum value	26,4	20,3	18,7	72,1	78,2
Minimum value	0,0	0,7	0,1	2,0	6,5

4. Writer Identification

As mentioned before, writer identification is the task related to determining the author of a sample handwriting from a set of writers [12]. Thus, Baranoski

et al. [1] performed an experiment based on relative slant as discriminative feature [2].

The feature extraction algorithm uses a structuring element on the image segment. The objective is to obtain L relative slant direction angles Θ. Depend on de structuring element we obtain a specific number of features L. The authors tested three different size of structuring element, with k = 3, 4, and 5 pixels and respectively L = 9, 13 and 17. The best result was obtained using k = 5, L = 17. The protocol used in the experiment considers that each known sample, belonging to the reference set (4 to 10 samples), is compared with the questioned or unknown writer sample. In this experiment a set of 5 reference samples was used for each writer. The best results applying SVM (Support Vector Machine) as classifier (kernel linear) achieved a false rejection rate of 1.73% and a false acceptance of 10.87%. These results demonstrated the discriminative capacity of the graphometric feature (relative slant) even being used in a global approach [1].

5. Conclusion and Future Works

In this paper we have introduced a forensic letter, the PUCPR letter, which is based on the Brazilian Portuguese language particularities. The database has 315 writers and we are continuing to collect more data. We have described different and important characteristics from the Brazilian writers about writing style, relative slant, relative relationship between letters and baseline, sentence indentation, and use of space. These characteristics are very often used by Forensic Document Examiners or in computer-based writer identification e.g. Wanda Measurement Tool [18] and other studies [3,8,19].

Our future works concern collecting more data and exploring other global and local features, such as relative spacing between words and between letters within the words, ascender, descender, and loops.

Acknowledgements

The authors wish to thank Cassiana Cunha Schepelski (Computer Science, PUCPR) which have collaborated to this work. This work has been supported by CNPq (grant 476637/2006-6 and grant 502338/2005-9).

References

[1] Baranoski, F.L.; Oliveira, L.S.; Justino, E.J.R.; Writer Identification Based on Forensic Science Approach. Conferencia Latinoamericana de Informática (CLEI2007), v. 1. pp. 25--32. 2007.

[2] Cha, S.H.; Use of the Distance Measures in Handwriting Analysis. Doctor Theses, State University of New York at Buffalo, EUA, 2001, p. 208.

[3] Cha, S.H.; Srihari, S.N.; Multiple Features Integration for Writer Identification. Seventh International Workshop on Frontiers in Handwriting Recognition (IWFHR2000), pp. 333—342. 2000.

[4] Franke, K. and Köppen, M.; A Computer-based System to Support Forensic Studies on Handwritten Documents. International Journal on Document Analysis and Recognition, 3(4), pp. 218--231, 2001.

[5] Gluhchev, G.; Handwriting in Forensic Investigations. Int. Journal Information Theories & Applications, Vol.11, pp. 42-46, 2004.

[6] Justino, E.J.R.; The Questioned Document Analysis (A Análise de Documentos Questionados), PUCPR, 2002.

[7] Koppenhaver, K.M.; Forensic Document Examination, Principles and Practices. Humana Press, 2007.

[8] Schlapbach, A.; Bunke, H.; Using HMM Based Recognizers for Writer Identification and Verification. 9th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR2004), pp. 167--172, 2004.

[9] Morris, N.; Forensic Handwriting Identification Fundamental Concepts and Principles, Academic Press, 2000.

[10] Oliveira, L S., Justino, E., Freitas, C.O.A., Sabourin, R.; The Graphology Applied to Signature Verification, 12th Conference of the International Graphonomics Society (IGS2005), pp. 286-290, 2005.

[11] Osborn, A.S.; Questioned document. 2nd ed. Albany, NY: Boyd Printing, 1929.

[12] R. Plamondon and G. Lorette. Automatic signature verification and writer identification – the state of the art. In Pattern Recognition, volume 22, pp. 107–131, 1989.

[13] Schomaker, L.; Advances in Writer identification and verification. 9th Int. Conf. on Document Analysis and Recognition (ICDAR2007), Key-Note speaker, 2007.

[14] Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.; Individuality of Handwriting: A Validation Study. 6th Int. Conf. on Document Analysis and Recognition (ICDAR2001), pp. 106-109, 2001.

[15] Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.; Individuality of Handwriting. In Journal of Forensic Sciences, Vol. 47(4), pp. 856–872, 2002.

[16] Srihari, S.N., Huang, C., Srinivasan, H.; Content-based Information Retrieval from Handwritten Documents. 1st International Workshop on Document Image Analysis for Libraries (DIAL2004), pp. 188-194, 2004.

[17] Tappert, C.C.; Suen, C.Y.; Wakahara, T.; The State of Art in On-line Handwriting Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, aug. pp.787-808, 1990.

[18] Van Erp, M., Vuurpijl, L., Franke, K., Schomaker, L.; The WANDA Measurement Tool for Forensic Document Examination. Journal of Forensic Document Examination, Vol. 16, pp. 103-118, 2004.

[19] Zhang, B., Srihari, S.N.; Binary Vector Dissimilarity Measures for Handwriting Identification. In Proc. SPIE, Document Recognition and Retrieval X, pp. 155-166, 2003.