Extracting Structural Features from Manuscript Texts for Forensic Documentoscopy and Graphoscopy Applications

Edgardo M. Felipe-Riverón, Salvador Godoy-Calderón and Edith C. Herrera-Luna

Centro de Investigación en Computación, Instituto Politécnico Nacional, Juan de Dios Bátiz and Miguel Othón de Mendizábal, P. O. 07738, Gustavo A Madero, México

edgardo@cic.ipn.mx,sgodoyc@cic.ipn.mx,edith.hluna@gmail.com

Abstract. This paper presents a new approach for extracting features from a manuscript, as well as a novel approach for modeling the graphoscopic structure of that manuscript at the word and the line levels by using the extracted features. As this new approach is independent of the document's semantics it allows the use of collective decision algorithms for author recognition tasks. Also, this approach represents a hybrid or eclectic paradigm between the texture-related and structure-related modeling approaches. The structural model is explained and the text features to be extracted are analyzed; then a series of test experiments in author identification, using the extracted features, is presented along with a comparison with other similar researches.

Keywords. Feature extraction, structural features, author identification, manuscript text, supervised classification.

1 Introduction

Nowadays, identification and verification are very common tasks that people face every day, not only when extreme security measures are needed, but as part of their daily routines within personal and professional activities. Identification and verification of a person is, nonetheless, a difficult task. That is why both theoretical and applied methods for these tasks are increasingly common, and a wide range of software and hardware tools, commercially available, allow these tasks to be performed with confidence, speed, and with a much smaller margin of error.

Graphoscopic techniques analyze a person's handwriting style and use that information for identification purposes, as well as to determine the specific circumstances under which each document was written. Under such consideration, Graphoscopy is both, a branch of Biometry [1] and a fundamental tool for Documentoscopy [14].

From the view point of Biometry, a person's writing style is a dynamic feature that, in combination with other biometric techniques, can be successfully used for identification purposes [2]. Criminalistics and Criminology research groups have developed a special interest on Documentoscopy techniques, not only for their

Advances in Computer Science and Engineering
Research in Computing Science 63, 2013 pp. 55–65
Paper Recived 17-09-2013 and Acepted 21-10-2013

potential capability to reveal the identity of a manuscript's author, but also for some other psychological traits about the author's health and frame of mind, it may reveal, by analyzing drawing samples, as well as handwriting and signature samples [3]. Other potentially interesting traits include, forfeit detection, multiple authorship identification, and historical edition profiling.

In order to analyze handwriting, two kinds of data acquisition techniques are traditionally used: online and offline. Online techniques acquire data directly from the author using some capturing device that delivers the data in real time and is capable of automatically identifying dynamic changes in the writing style, such as changes in pressure, speed, and inclination. On the other hand, offline techniques acquire data by analyzing texts previously written on some support surface (paper, parchment, leather, etc.). This type of analysis is usually performed by extracting features from digital images of the text, and thus, it is more suited for forensic applications.

The main goal of this work is to propose a particular approach for modeling a manuscript, as well as the particular set of features that need to be extracted from a digital image of the manuscript. Also, we would like to show how the proposed set can be effectively used for forensic Documentoscopy and Graphoscopy applications, with sample results from a previous author identification research. The proposed model effectively constitutes a different and hybrid approach between the classic texture-related and the structure-related approaches to handwritten text's feature extraction.

2 Background

One of the most relevant researches concerning the writing of an individual as a recognizable unique mark of each person was performed by Sargur N. Srihari *et al.* [5, 6]. Srihari's work group created a database with writing samples of approximately 1,500 participants, who were asked to write a letter in English, over a blank sheet of paper [4, 6]. This research identified and analyzed text features from two groups: conventional and computational features. Conventional features are found by the manual analysis of graphoscopic specialists, while computational features come from the analysis of some intermediate media (typically a digital image), and extracted with morphologic operators and/or data transformations like the Fast Fourier Transformation or Wavelets [14]. The set of computational features extracted, was divided in two groups, called macro-features and micro-features. Macro-features are extracted at the whole document level, as well as at the paragraph, row and word levels of analysis. Micro-features, on the other hand, are extracted at the character and stroke levels. For testing purposes, they selected a subset of 11 macro-features, and 3 micro-features, and reported an effectiveness range from 78% to 96% when using those features for author identification. In [7], H.E. Said *et al,* proposed an algorithm for offline data analysis, based on texture features. By using multi-channel filters, Grey Scale Co-occurrence Matrices (GSCMs), and a K-Nearest Neighbor with weighted Euclidian distance criteria as a classifier, they achieved a 96% rate of correct author identification, over a database of 150 writing samples. The main relevance of this research lies on the fact that all the extracted features can be used in independence from the document's content, and its text distribution. E.Zois and V.

Anastassopoulos, proposed in [9], a distinct algorithm that uses horizontal projections and morphologic operators, as well as a texture-related analysis over English and Greek handwritten words. To prove their method to be language independent, they constructed a sample database from 50 authors, each writing 45 different words in English and Greek. In general terms, their method applies a threshold over each word's image, then, a series of morphological thinning and opening operations. A combination of multi-layer perceptron and Bayesian analysis is used to classify samples, and they report a 90% precision in identifying authors. In [1], a recognition method, using some of the features proposed by Srihari *et al.*, was shown, although they were extracted at the character level exclusively. The proposed method starts form digital images, scanned at 300 dpi, where all the characters are manually segmented and saved into a database. After some preprocessing to get each character in a binary form, all GSC [8, 9], and geometric features [10] are automatically extracted, as well as some novel gradient direction related features. At the end of this research, authors report that, using a K-Nearest Neighbor classifier, gradient features better allow them to reach a 100% precision in identification tasks, while the whole set of the same features, along all GSC, and geometric features, only allows them a 90% precision. All their reported experiments were performed over a 1,400 characters database, with 40 different authors. Along the same line of thought, on processing individual characters, is the research performed by V. Pervouchine in [11]. His work focus on the specific analysis of letters 'd', 'y', and 'f', as well as the 'th' combination in English texts. After constructing a custom database, with 15 to 30 different writing samples, from 150 authors, Pervouchine managed to extract a total of 31 features, and selected 13 of them to be essential, 14 to have only partial relevance, and the last 4 to be irrelevant. However, his reported results show an identification performance of only 58%. Finally, a generally accepted top reference in the field is the work of Bansefia *et al.* in [12]. This workgroup used two different databases; the first one, custom-made by them, was called *PSI-Database*, and contained a sample handwritten letter, with 107 words, from 88 French authors. The second database was the relatively famous *IAM* database, which includes samples from approximately 150 authors, and is freely available on [13]. Their proposed processing technique, separates the text in graphemes, based on the analysis of the upper outline of the letters, and they report a 95% precision identifying authors from the *PSI-DataBase*, while only 86% with authors from the *IAM* database.

3 Handwriting characteristics

In order to analyze a handwriting sample, several distinct measurements are required. The measured elements of a person's writing form two disjoint groups, generally called formal elements and structural elements. See Figures 1 and 2.

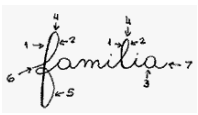


Fig. 1. Formal elements of handwritten text: (1, 2, 4) Crest, (3) Oval, (5) Axis, (6) Initial & (7) Final point.

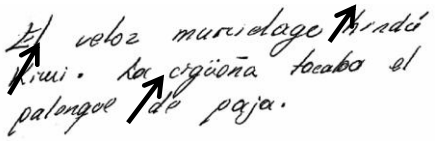


Fig. 2. Example of the structural element called **Slant**. The text in this figure has a right slant.

Formal elements are those characteristics that form letters or words, like strokes, lines, initial and end points, etc. Also, as described in [14], the closed central portion of letters, called ovals, can also be considered as formal elements. The superior portions of letters, called crests, and the axis that guide the lower portion of letters, as well as the punctuation signs like colons, semicolons, etc., are considered formal elements as well. In contrast, structure elements are those characterizing the authors writing style, such as its size, form, direction, and organization. Although both groups of elements have been studied for the last 100 years, there is still no agreement on a unified set of features that should be considered.

The measurement of formal and structural features requires the definition of writing zones which are defined with the aid of two imaginary horizontal lines, tangent to the upper and lower parts of the text, and which delimit three writing zones. The middle zone includes most of the writing strokes, the upper zone includes the letters crests, and finally, the lower zone includes the axis of letters (Figure 3).

Fig. 3. Writing zones (1) upper zone, (2) middle zone, and (3) lower zone.

The most relevant handwritten text features for this research follows:
Order (Regularity). It refers to the distribution and organization among letters, words, text lines, paragraphs and text margins. **Size (Dimension).** Refers to the breadth of writing, and the amount of space each letter occupies within a word, or a word within a line. The proportion among writing zones, as well as the change in size from lower letters to capital letters is also considered (See [3, 15]). **Proportion.** This can be modeled as a subset of the size measurements, or it can be based on the size ratio of the writing zones, the ratio of the document's margins, or the ratio between the area occupied by the text and the whole area available on the writing media. **Shape.** Measures some traits of the writing, such as if letters are angled, curved, typographical or decorated (Widely in [3, 15, 18]). **Angularity.** This trait can well be a subset of the shape measurements. The text angularity refers to the presence of angles and curves in the writing, the angled end of each character or word, as well as the breadth of the curved strokes. **Direction.** The trajectory followed by the lines or strokes that comprise characters. Some researchers consider the slope of the writing zone, while others prefer the angle of an imaginary line below the text (Refer also to [2, 3]). **Slant. It r**efers to the angle of single words, lines, or paragraphs. The standard is to look for right-slant or left-slant. **Continuity and Linkage.** It is the measure of the strokes that bind two consecutive words in a line of text. **Separation and Cuts**. It measures the blank space between letters, words or text lines. These gaps are also considered as writing strokes.

In order to measure all the above traits, and construct an adequate database for offline analysis, test subjects were asked to write over white Bond paper and with the same writing tool. The complete set of details about the database and the form filled by test subjects can be found in [15].

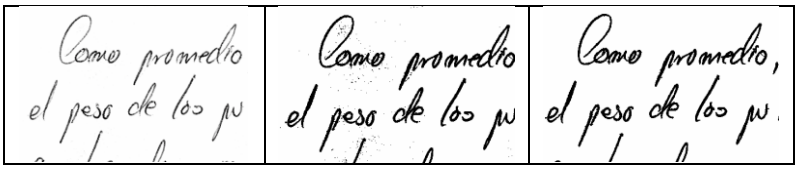
The digital image of the manuscript is thresholded to get a binary image, and then the horizontal and vertical projections of the binary image are used. For the horizontal projection, the profile of each row of pixels is considered, and the amount of those pixels with a zero value (pixels in black) is used to construct a histogram. The vertical projection follows the exact same procedure, but using the profile of each column of pixels in the image. A detailed description about this can be found in [15, 18].

The rest of the macro-features were extracted by using morphological dilations and erosions, with squared and straight line structure elements previously rotated. Other morphological operators, such as openings, closures and geodetic reconstructions were also used. To review these procedures refer to [15, 19].

4 Methodology

Our general methodology starts with a data acquisition phase, scanning the test forms with a standard color scanner at 300 dpi, and setting up a database with them. Before the preprocessing phase takes place, the characteristic writing zones are calculated for each scanned form. Then, after enhancing the sample images, the handwritten text is segmented and features at the word, line and paragraph are extracted. Once a full supervision sample has been constructed, the specific comparison criteria are selected and each extracted feature is weighted. Some important details about the procedure follow.

During the preprocessing phase, two different methods are used to enhance the digital image of each text. First, a threshold is applied over the green plane, with the Otsu method [16]. Also, the same original green plane is processed using the Khashman and Sekeroglu algorithm [17]. The result from the Otsu method turns out to be very clear; however it removes some important regions from the manuscript. On the other hand, the Khashman and Sederoglu algorithm yields a very noisy image, but the text is much better shown in it. Therefore, a geodetic reconstruction, following the morphological operation of erosion, is used. The image resulting from the Khashman algorithm is used as a mask, and the Otsu image is used as the mark for this reconstruction procedure. As a result, we get a very clean image, without noise at all, but with the handwritten text shown as best as possible (Fig. 4)



A B C

Fig. 4. (A) is the original image, (B) is the result from the Khashman and Sekeroglu algorithm, and (C) is the geodetically reconstructed image.

The exact set of extracted features is the following:

- Margin proportionality features
  1) Average paragraph's distance to the left and right physical margins (R_1, R_2).
  2) Average text line's distance to the left and right physical margins (R_3, R_4).
- Direction and size features
  3) Average distance between successive rows (R_5).
  4) Ratio of the space occupied by a row and its distance from the previous and the following rows (R_6, R_7).
  5) Average space among words from a same row (R_8).
  6) Average row direction (R_9).
- Writing zone's features
  7) Upper zone – Middle zone ratio (R_{10}).
  8) Lower zone – Middle zone ratio (R_{11}).
- Slant features
  9) Average word slant (R_{12}).

5 Classification model

The margin proportionality feature is extracted at the paragraph and line levels, since different indentation formats can be notoriously discriminant under some circumstances. In both cases, the minimum size rectangle, containing the paragraph or row, is cut from the binary image, then, we simply count the number of pixels, form the original image, that were cut, so extracting the first two features R_1 and R_2 (Fig. 5).The same minimum size containing rectangle procedure is applied at the text line level, and the comparison between successive rectangles yields features R_3 and R_4 (Fig. 6).

The first six features are measured directly from the text lines on the manuscript. The next features are calculated as an average of some other features at the word level. Words with crest or axis are clustered into four groups: words with crest, without crest, word with axis, and without axis. For each group features R_{10}, R_{11} and R_{12} are calculated. These features are extracted at the paragraph, line, and word levels. However, since a feature vector represents a text line, features R_1, R_2, and R_5, which are paragraph features, are not included. Also, these three features are equivalent, at the line level, to features R_3, R_4, and R_6.

Then, each pattern represents a line of text in the document, and is composed by 22 features, calculated with the text traits extracted previously. From the form filled by test subjects, 605 text lines, from 30 different authors, were obtained.

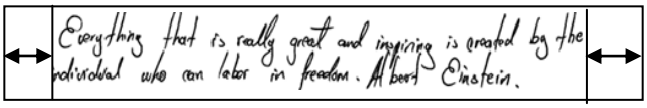


Fig. 5. Features R_1 and R_2.

Fig. 6. Features R_3 and R_4.

5.1 Feature selection

In order to compare patterns, a similarity or dissimilarity function must be used. A weighted syntactic distance is selected for that task. The idea is to allow automatic systems the opportunity to assign a greater weight or relevance to the most typical traits in the class that describes the writing of a person. A complete description of the weighting scheme can be found in [29, 30 and 31].

A collective decision algorithm is used to identify the author of a manuscript by means of the individual classification of the text lines that comprise the manuscript. The descriptive features used to make up the patterns representing such lines include features extracted at the word level, as well as at the line level. When all the text lines have been classified, a final collective decision regarding the author of such text is applied.

6 Experimental results and precision assessment

An *ad hoc* database was created with manuscripts written by 50 test subjects, who wrote three handwritten texts. Each manuscript contains from 5 to 9 text lines, giving closely a total of 600 lines. Text contents were selected arbitrarily from non-technical books with no restrictions on the logic or semantics. Images of those manuscripts were digitally scanned at 300dpi with a conventional scanner and all manuscripts were written with the same black ink pen over white paper.

Three different types of experiments were performed: in experiments of type1 the supervision sample contains the most representative patterns in each class (e.g. those patterns with the maximum average similarity with all others in the same class); in experiments of type2 it contains a random selection from all the patterns in all classes; and finally, in experiments of type3 the supervision sample contains only the least representative patterns from each class (e.g. those with the minimum average similarity with all others in the same class).

Three class representative patterns were selected from each class within each supervision sample, according to the previously described procedure. Experiment results are shown in Table 1.

In the table, the use of a differentiated weighting scheme significantly increases the classification rate, both for text lines and for the whole manuscript. Turns out that this type of weighting allows a more precise characterization of an author's writing style. For experiments of type 2, the higher classification rate is 88.89%.

Table 1. Manuscript line and whole text classification rates for experiments of type 1, 2 and 3.

No.	Text line classification rate (%)	Manuscript classification rate (%)
Experiment type 1		
1.1	42.28	61.11
1.2	56.10	66.67
1.3	60.16	72.22
1.4	56.10	77.78
1.5	73.17	94.44
1.6	47.97	66.67
1.7	61.79	72.22
1.8	60.16	63.89
1.9	51.22	72.22
1.10	63.42	83.33

No.	Text line classification rate (%)	Manuscript classification rate (%)
Experiment type 2		
1.1	52.76	75.00
1.2	67.72	88.89
1.3	64.57	94.44
1.4	58.26	83.33
1.5	70.08	88.89
Experiment type 3		
1.1	46.46	66.67
1.2	65.35	88.89
1.3	62.99	83.33
1.4	58.26	83.33
1.5	72.44	94.44

A comparative analysis with several state-of-the-art works in manuscripts author identification shows some advantages of the method herein proposed. First, there is the opportunity to increase the classification rate, just by changing the weighting scheme. Second, there is also a chance to change the function for comparing text patterns. Third, although a big number of text line patterns can eventually lead to a deeper analysis of a manuscript, usually a small number of text lines can be enough for achieving high classification rates.

Text independence is generally a difficult problem when automatically analyzing manuscripts. However, results obtained by this research are not significantly different from those obtained by other research groups. Table 2 compares the classification rate of this work with that obtained by several other relevant studies in the field, as presented by [15].

Although each research uses a different sample database, the relative complexity for feature extraction of the database used for this research can be regarded as slightly higher as that from all other works, since those other works usually show a bigger inter-line spacing which simplifies the text segmentation procedure.

Table 2. Comparison with classification rate results in other researches.

	# of Authors	Samples	Text dependency	Classification rate (%)
Said et al. [8]	2x20	25 blocks of text	NO	95
Zois y Anastassopoulos [9]	50	45 samples of the same word	YES	92.48
Marti and Bunke [13]	20	5 samples of the same text	YES	90
A. Bensefia et al. [12]	88	Paragraphs of 3-4 words	YES	93 / 90
This work	**30**	**3 samples of the same text**	**NO**	**94.44**

7 Conclusions

Two clearly distinct methodological phases can be identified in this proposal: the first one deal exclusively with feature extraction, the second one, does characterization and classification. These two phases are not strictly sequential, however. As the first text features are extracted, an adequate similarity/dissimilarity measure[1] is selected, as well as all related comparison criteria. Semantics of each text trait must be carefully analyzed, in order to set an appropriate mixing of features. The weighting scheme turns out to be extremely useful when characterizing an author's handwriting style.

A comparison with previously published works shows that the modeling approach herein proposed yields better results, with the added advantage that the recognition process needs not to be dependent on the semantic contents of the text. The implementation of these improvements may be extremely useful for forensic Documentoscopy and Graphoscopy applications, as well as for more traditional authentication and security systems.

Acknowledgements

The authors of this paper wish to thank the Centro de Investigación en Computación (CIC), Mexico; Research and Postgraduate Secretary (SIP), Mexico, and Instituto Politécnico Nacional (IPN), Mexico, and CONACyT, Mexico, for their economic support to this research.

References

1. Tapiados Mateos, Marino, Sigüenza Pizarro, Juan A. Tecnologías biométricas aplicadas a la seguridad. 1a. México, D.F.: Alfaomega Grupo Editor, 2005. ISBN: 970-15-1128-X.
2. Del Val Latierro, Félix. Grafocrítica. El documento, la escritura y su proyección forense. Madrid, España: Tecnos S.A., 1963.
3. Tesouro De Grosso, Susana. Grafología. Análisis e interpretación científica de la escritura. Buenos Aires, Argentina: Editorial Kier, 2006. ISBN: 950- 17-7011-7.
4. Moreno, G. R. y García, R. Temas de Criminalística. Jornadas sobre Justicia Penal sobre Temas de Derecho Penal, Seguridad Pública y Criminalística. s.l.: Universidad Nacional Autónoma de México, 2005.
5. Srihari Sargur, N. Handwriting identification: research to study validity of individuality of handwriting and develop computer-assisted procedures for comparing handwriting. University of Buffalo, U.S.A.: Center of Excellence for Document Analysis and Recognition, 2001. Tech. Rep. CEDAR-TR-01-1.
6. Srihari, Sargur, N. et al. Individuality of Handwriting. E.U.: Journal of Forensic Sciences, 2002. Vols. 47, No. 4. Paper ID JFS2001227-474.
7. Said, H., Tan, T. y Baker, K. Personal Identification Based on Handwriting. s.l.: Pattern Recognition, 2000. págs. 149 -160. Vol. 33 No. 1.

[1] Although in theory there are many similarity/dissimilarity functions that can be selected, the authors have found that a term-by-term comparison on the patterns (a syntactic distance) usually promotes faster classification processes, as well as clearer interpretations of the obtained results.

8. Said H. E., et al. Writer identification from non-uniformly skewed handwriting images. s.l.: in Proc. 9th British Machine Vision Conference, 1998. págs. 478-487.
9. Zois, E y Anastassopoulos, V. Morphological waveform coding for writer identification. s.l.: Pattern Recognition, 2000. págs. 385 - 398. Vol. 33. No. 3.
10. Srihari, Sargur, N. Recognition of handwritten and machine-printed text for postal address interpretation. s.l.: Pattern Recognition Letters, 1993. págs. 291-302. Vols. 14, No. 4. ISSN: 0167-8655.
11. Pervouchine, Vladimir y Leedham, Graham. Extraction and analysis of forensic document examiner features used for writer identification. s.l.: Pattern Recognition, 2007. págs. 1004-1013. Vol. 40. ISSN: 0031-3203.
12. Bensefia A., Paquet, T. y Heutte, L. A writer identification and verification system. s.l.: Pattern Recognition Letters, 2004. Vol. 26. 2080–2092.
13. Bunke, H. y Marti, U. A full English sentence database for off-line handwriting recognition. [ed.] University of Bern Institute of Informatics and Applied Mathematics. s.l.: Proc. of the 5th Int. Conf. on Document Analysis and Recognition, ICDAR´99, 1999, pp. 705-708. http://www.iam.unibe.ch/fki/databases/iam-handwriting-database/iamhandwriting- database#icdar99.
14. Del Picchia, José y Celso. Tratado de documentoscopía: La falsedad documental. Argentina: Ediciones La Rocca, 1993. ISBN 959-97-1450-4.
15. Guzmán, Carlos Alberto. El peritaje caligráfico. 1a. y 2a. reimp. Buenos Aires: Ediciones La Rocca, 2005. pág. 336. ISBN: 950-9714-59-3.
16. Otsu N., A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man and Cybernetics, 9(1): pp. 62-66. (1979).
17. Khashman, Adnan y Sekeroglu, Boran. A Novel Thresholding Method for Text Separation and Document Enhancement. Near East Univ., Mersin: ICIT 2006. IEEE International Conference on Industrial Technology, 2006. ISBN: 1-4244-0726-5.
18. Herrera-Luna, Edith C., Identificación del autor de un texto manuscrito., Tesis de Maestría en Ciencias de la Computación, Centro de Investigación en Computación, Instituto Politécnico Nacional. México D. F., México.
19. Soille, Pierre. Morphological Image Analysis. Segunda Edición. s.l.: Springer. pág. 391. ISBN: 3-540-42988-3.