Multi-level Modeling of Manuscripts for Authorship Identification with Collective Decision Systems

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

Center for Computing Research, National Polytechnic Institute,
Juan de Dios Bátiz and Miguel Othón de Mendizábal, P.O. 07738, Gustavo A Madero, México
{sgodoyc,edgardo}@cic.ipn.mx, edith.hluna@gmail.com

Abstract. In the context of forensic and criminalistics studies the problem of identifying the author of a manuscript is generally expressed as a supervised-classification problem. In this paper a new approach for modeling a manuscript at the word and text line levels is presented. This new approach introduces an eclectic paradigm between texture-related and structure-related modeling approaches. Compared to previously published works, the proposed method significantly reduces the number and complexity of the text-features to be extracted from the text. Extensive experimentation with the proposed model shows it to be faster and easier to implement than other models, making it ideal for extensive use in forensic and criminalistics studies.

Keywords: Collective decision, Author identification, Manuscript text, Supervised pattern recognition.

1 Introduction

The analysis of handwritten text, for the purpose of authentication and author identification, is a tool that allows researchers to study text from many different points of view according to the number of authors, the type and quantity of characteristics extracted from the text, the classification algorithms used for its recognition, and so on. The analysis of manuscripts is at the core of graphoscopic analysis techniques and plays an important role in a number of practical problems, included, but not limited to, forensic and criminalist processing of evidence, validation of legal documents, historiography, psychological profiling, etc.

Since the 1950's several manuscript author identification and verification methods have been developed for forensic document analysis. Before the era of digital communication, the wide range of implements and supports for writing motivated several approaches to characterize handwriting. Those approaches differ among themselves by their way of capturing data and by their dependence on the semantics of the written text.

Computer aided manuscript analysis is a broad field of study that encompasses two main areas: Optical Character Recognition (OCR) and Writing Identification and

* Corresponding author.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 757–764, 2012.
© Springer-Verlag Berlin Heidelberg 2012

Verification (WIV). OCR has been widely studied and consists of recognizing characters from a digital image of a manuscript in such a way that the text may be interpreted word by word or even symbol by symbol [19].

The WIV area deals with the recognition of a manuscript's author, the relationship of several authors to different documents, the identification of document alterations, etc. In some WIV approaches it is necessary to know the semantic contents of the words that make-up the text (text-dependent methods) [20], and in others, the method does not depend on semantic contents (text-independent methods) [2].

Computer-related characteristics extracted from the text may be texture-related or structure-related [3]. When extracting texture characteristics, the document is seen and processed as an image (not as a text), but when structural characteristics are extracted, a description similar to that given by a graphoscopic expert is sought in order to characterize the properties of the author's writing style.

This paper introduces a new approach to modeling handwritten text by extracting off-line static features from the manuscript at the line and word levels. This method allows the use of collective decision algorithms for author recognition tasks. The proposed approach is independent of the semantics and represents a hybrid or eclectic approach between texture and structural characteristics.

2 Related Works

Recently, amongst already published papers are [1] and [4], where the identification efficiency is considerably reduced due to the high number of authors in the supervision sample. The paper [5] propose a manuscript's texture analysis technique using Gabor's filters and grey-scale co-occurrence matrices (GSCM) using the weighted Euclidean distance and a K-NN classifier [6] to identify 40 different authors. Authors in [7] propose the use of horizontal projections and morphological operators together with the texture characteristics of English and Greek words, using a multilayer perceptron and a Bayesian classifier.

The method proposed in [3] splits words into graphemes and at the same time combine the local characteristics extracted from regions of text. Finally they employ a commonly used model for information retrieval, known as vector space model, for the task of identifying the author of a manuscript. The research outlined in [8] presents an algorithm that extracts structural features from characters and graphemes, by means of a genetic algorithm, to look for optimal characteristics after extracting 31 characteristics from each selected word and using a neural network classifier. In their doctoral dissertation [2] extract textural features from manuscripts and then use probability functions and an Euclidean and Hamming distance-based K-NN classifier to identify and verify the author. In [9] is extracted a set of structural features from text lines, as well as a set of fractal-based features, resulting in a 90% plus efficiency of author recognition.

In [10] a study can be found about handwritten text recognition, taking into account methodologies for working in-line and off-line. In [11] the IAM database is described in a general way and an analysis is made about the way in which images included were segmented. The last two papers include references that detail recent research work on manuscript analysis.

More recently, in [17] a segmentation methodology, in text lines and words, of handwritten documents is presented. Text line segmentation is achieved by applying Hough transform on a subset of the document image connected components. With careful post-processing steps authors claim to achieve an efficient separation of vertically connected characters using a novel method based on skeletonization. They use several performance measures to compare the text line segmentation and word segmentation results against the available supervision sample. Paper [19] shows a methodology for off-line handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images have an approximately equal number of foreground pixels. Feature extraction is followed by a two-stage classification scheme based on the level of granularity of the feature extraction method.

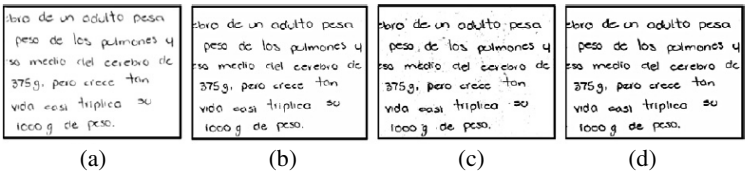
Until today the most reliable results achieved in authorship recognition over manuscripts can be found in [3], [4], [5], [8], [12] and [17].

3 Proposed Model

The basic idea behind the proposed model is to extract from the manuscript a different set of features at the line and word levels. The final model uses patterns that describe the manuscript at the paragraph level by selecting and averaging all features extracted from lines that comprise each paragraph. Extracted features are then processed at different stages by an algorithm that assigns a potential author to each paragraph and then decides the authorship of the whole manuscript by collecting all the evidence learned from each paragraph in the text.

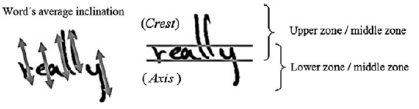
Manuscript images are first pre-processed in order to sharpen the digital image and remove noise. A binary image of the manuscript is obtained by thresholding over the green plane of the original color image with Otsu [13] and Khashman - Sekeroglu [14] methods. Finally, a geodesic reconstruction of each manuscript is obtained by erosion using the Otsu-processed image as a marker (see Figure 1). All features are extracted from the binary image of each manuscript.

At the line level the space percentage occupied by the left margin, the right margin, the separation between subsequent lines, the general direction of the writing and the inter-word space are considered. Features extracted at the word level include the proportions of the middle zone of the writing compared to that of the upper and lower zones, word inclination and the presence of a crest and an axis in all words. Figure 2 shows the semantics of all features extracted at the line and word levels.

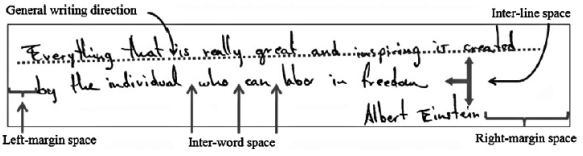


(a) (b) (c) (d)

Fig. 1. (a) Sample manuscript color image, (b) Otsu-thresholding over the green plane, (c) Khashman – Sekeroglu thresholding, (d) final noise-free binary image



(a) Features at the word-level



(b) Features at the line-level

Fig. 2. Features extracted at different levels

Each manuscript's text line is represented by a 22-tuple with word-level and line-level features. Word-level features are extracted from words with and without upper-zone, as well as from words with and without lower-zone. All features are positive real or integer numbers, except for the Average inclination of words, which takes values in the interval [-1, 1] of real numbers. Nevertheless a good recognition method must be independent of the representation space for the objects. Table 1 sums up the features extracted at the word and line levels.

Table 1. Features representing a line of text

Word-level features	Line-level features
Average upper/middle zone ratio	Left and right margins space percentage
Average lower/middle zone ratio	Current/previous line ratio
Average inclination of words	Current/next line ratio
Number of words in the line	Average inter-word space
	General direction of writing.

4 Recognition Process

The manuscript whose author is to be recognized is pre-processed and the aforementioned features are extracted to form several line and paragraph patterns and a weighted syntactic-distance function is used for comparison between patterns. The particular differentiated-weighting scheme used assigns a different feature weight value depending on the class being tested. This mechanism provides enough flexibility to accurately discriminate patterns belonging to classes where the same sub-set of features is relevant but with a different proportion in each case. Details about such a scheme and its use for author recognition can be found on [15].

Each text-line pattern is independently classified and then a collective decision rule is applied, issuing a final decision regarding the authorship of the text. This final decision may be one of the authors included in the supervision sample or an unknown author not included in the sample, if certain similarity thresholds previously established are not met. Pseudo code of the specific procedure looks as follows:

1. For each pattern in the supervision sample, determine its class representativeness.
2. Select only the most representative patterns to form a Reduced Supervision Sample.
3. Compare each control pattern with the reduced supervision sample formed in step 2.
4. Calculate the average similarities of control patterns vs representative patterns.
5. Use the highest average similarity as the final decision rule.
6. If considered convenient, add the recently identified patterns to the supervision sample and recalculate the set of objects that are representative of each class.

The classification of the manuscript is a collective decision based on the lines contained in a given text. The procedure for its classification looks as follows:

1. Create a pattern for each text-line in the manuscript and classify it.
2. The whole manuscript is labeled as the class which contains the majority of its line-patterns, allowing a maximum of classes.

This multiple-pattern representation of each manuscript, along with the collective decision criteria used by the solution rule, establishes a new and not previously explored approach for this kind of problems. The resulting impact over the precision and general efficiency of the identification process can be seen on the next section.

5 Experimental Results

An *ad hoc* database was created with manuscripts written by 50 test subjects. Each subjected wrote three handwritten texts, always using print-type letters. 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 and white paper.

Three supervision and control samples were built for the experiments to take place. For experiment type #1, the supervision sample contains the most representative patterns in each class. In experiment type #2 the total number of manuscripts is randomly divided between the supervision sample and the control sample. Finally, for experiment type #3 only the least representative patterns from each class was selected for the supervision sample. In each experiment type the objects from the control sample are classified taking the supervision sample objects as a reference. Three class-representative patterns were selected from each class within the supervision sample, according to the previously described procedure.

The following rule was used for assessing the efficiency of the manuscript classification: identification is considered correct if and only if the manuscript is classified in less than q classes and one of them is the correct one. Table 2 shows some experimental results obtained, for all three experiments types, when using a differentiated feature-weighting scheme.

Table 1. Results from the first experiment set

Experiment type	% of text-lines recognition	% of manuscript recognition
1	60.16	72.22
1	61.79	72.22
2	67.72	88.89
2	64.57	94.44
3	65.35	88.89
3	62.99	83.33

Although the percentage of correct text-line recognition seems unacceptably low, results only got better for the manuscript level, reaching levels higher than 80% in several cases. Although the classification at the line level was substantially altered, the impact of the differentiated feature-weighting scheme becomes evident in the results at a text level.

A second set of experiments was carried out, considering the centroid of each class as the only representative pattern for that same class. This raised the effectiveness of the recognition for the line level and to a lesser degree for the classification of the whole manuscript (See Table 3).

Table 4 shows this research's results compared with those by other authors (taken from Bensefia et al. 2005). The comparison is summarized and some details are added on the type of methodology applied.

Table 2. Results from the second experiment set

Experiment type	% of text-lines recognition	% of manuscript recognition
1	73.17	94.44
2	70.08	88.89
3	72.44	94.44

Table 3. Comparison of manuscript's author identification

Publication	Number of writers	Supervision Sample	Lexicon dependency	Reported Performance (%)
Said et al. (2000) [5]	40	Few lines handwritten text	No	95.0
Zois and Anastassopoulos (2000) [7]	50	45 examples, the same word	Yes	92.48
Marti et al. (2001) [16]	20	5 examples of the same text	Yes	90.00
Bensefia et al. (2005)[3]	88	Paragraphs / 3-4 words	No	93.0 / 90.0
This paper	**30**	**3 examples of the same text**	**No**	**94.44**

6 Conclusions

A collective decision algorithm with a differentiated feature-weighting scheme 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. The highest efficiency percentages on identification are achieved in those experiments in which centroids are used as class representative patterns.

A comparison with previously published works shows that the modeling approach herein proposed yields better results than previous related works, 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 the identification of authors of handwritten texts, mainly in forensic control situations as well as in authentication and security institutions.

Acknowledgements. The authors would like to thank the Academic Secretary, COFAA, Postgraduate and Research Secretary, and Centre for Computing Research of the National Polytechnic Institute (IPN), CONACyT and SNI, for their economic support to carry out this work.

References

1. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of Handwriting. Journal of Forensic Sciences 47(4), Paper ID JFS2001227-474 (2001)
2. Pecharromán-Balbás, S.: Reconocimiento de escritor independiente de texto basado en características de textura. Tesis doctoral. Escuela Politécnica Superior, Universidad Autónoma de Madrid (2007)
3. Bensefia, A., Paquet, T., Heutte, L.: A writer identification and verification system. Pattern Recognition Letters 26, 2080–2092 (2005)
4. Srihari, S.N.: Recognition of handwritten and machine-printed text for postal address interpretation. Pattern Recognition Letters 14(4), 291–302 (1993)
5. Said, H., Tan, T., Baker, K.: Personal Identification Based on Handwriting. Pattern Recognition 33(1), 149–160 (2000)
6. Cover, T.M., Hart, P.E.: Nearest neighbour pattern classification. IEEE Trans. Inform. Theory, IT-13(1), 21–27 (1967)
7. Zois, E., Anastassopoulos, V.: Morphological waveform coding for writer identification. Pattern Recognition 33(3), 385–398 (2000)
8. Pervouchine, V., Leedham, G.: Extraction and analysis of forensic document examiner features used for writer identification. Pattern Recognition 40, 1004–1013 (2007)
9. Hertel, C., Bunke, H.: A Set of Novel Features for Writer Identification. In: Proc. Fourth Int'l Conf. Audio and Video-Based Biometric Person Authentication, pp. 679–687 (2003)
10. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: A comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 63–84 (2000)
11. Zimmermann, M., Bunke, H.: Automatic segmentation of the IAM off-line handwritten {English} text database. In: 16th International Conf. on Pattern Recognition, Canada, vol. 4, pp. 35–39 (2002)

12. Srihari, S.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. Tech. Rep. CEDAR-TR-01-1 (2001)
13. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Sys., Man., Cyber. 9, 62–66 (1979)
14. Khashman, A., Sekeroglu, B.: A Novel Thresholding Method for Text Separation and Document Enhancement. In: Proceedings of the 11th Pan-Hellenic Conference in Informatics, Greece, pp. 324–330 (2007)
15. Herrera-Luna, E., Felipe-Riverón, E., Godoy-Calderón, S.: A supervised algorithm with a new differentiated-weighting scheme for identifying the author of a handwritten text. Pattern Recognition Letters 32(2), 1139–1144 (2011)
16. Marti, U.V., Messerli, R., Bunke, H.: Writer identification using text line based features. In: Proc. ICDAR 2001, pp. 101–105 (2001)
17. Louloudis, G., Gatos, B., Pratikakis, I., Halatsis, C.: Text line and word segmentation of handwritten documents. Pattern Recognition 42, 3169–3183 (2009)
18. Bertolami, R., Bunke, H.: Hidden Markov model-based ensemble methods for offline handwritten text line recognition. Pattern Recognition 41, 3452–3460 (2008)
19. Vamvakas, G., Gatos, B., Perantonis, S.J.: Handwritten character recognition through two-stage foreground sub-sampling. Pattern Recognition 43, 2807–2816 (2010)
20. Jou, C., Lee, H.C.: Handwritten numeral recognition based on simplified structural classification and fuzzy memberships. Expert Systems with Applications 36, 11858–11863 (2009)