Machine-Printed from Handwritten Text Discrimination

Ergina Kavallieratou
TEI of Ionian Islands
ergina@wcl.ee.upatras.gr

Stathis Stamatatos
TEI of Ionian Islands
stamatat@teiion.gr

Hera Antonopoulou
Computer Technology Institute
antonopl@cti.gr

Abstract

This paper deals with the discrimination between machine-printed and handwritten text, a prerequisite for many OCR applications. An easy-to-follow approach is proposed based on an integrated system able to localize text areas and split them in text-lines. A set of simple structural characteristics that capture the differences between machine-printed and handwritten text-lines is presented and preliminary experiments on document images taken from databases of different languages and charcateristics show a remarkable performance.

1. Introduction

The presence of printed and handwritten text in the same document image is an important obstacle towards the automation of the optical character recognition procedure. Machine-printed and handwritten text should be processed using different methods and/or parameters in order to optimize recognition accuracy. Hence, integrated systems of document image analysis would benefit from classifying text in printed and handwritten areas before further processing.

Both machine-printed and handwritten text are often met in application forms, question papers, mail as well as notes, corrections and instructions in printed documents. In all mentioned cases it is crucial to detect, distinguish and process differently the areas of handwritten and printed text for obvious reasons such as: (a) retrieval of important information (e.g., identification of handwritten names in application forms), (b) removal of unnecessary information (e.g., removal of handwritten notes from official documents), and (c) application of different recognition algorithms in each case.

Previous work on this subject deals with the classification of text areas on the line-level, word-level or character-level, for Latin , non-Latin, or bilingual documents. Zheng et al. [1] perform text identification in noisy documents with comparative results for all levels. Fan et al. [2] perform detection of handwriting using structural characteristics for Chinese and English and report an accuracy rate of 85%. Pal et al. [3] process Indian scripts and the reported accuracy rate reaches 98.6%. Nitz et al. [4] apply text detection for mail facing and orientation purposes but no accuracy rate is mentioned for this specific task. Ma et al. [5] localize non-Latin script in Latin documents.

In this paper, we propose an easy-to-follow trainable approach to identify machine-printed and handwritten text areas. To this end, an integrated system able to localize text areas and split them into text-lines is used. In order to capture the differences between machine-printed and handwritten text-lines we introduce a set of simple and easy-to-compute structural characteristics. Experiments on document images taken from IAM-DB [6] and GRUHD [7] databases, of English and Greek respectively, are presented in order to examine the performance of the proposed approach.

This paper is organized as follows: In section 2 the overall system is presented emphasizing on the feature extraction procedure. Section 3 includes the evaluation experiments and section 4 summarizes the conclusions drawn from this study and future work directions are given.

2. The proposed apprach

The presented system processes a document image based on three main stages: i) the *preprocessing* module where, for a given document image, a set of text areas is localized resulting a series of text-lines, ii) the *feature extraction* module where a vector of structural characteristics that represents the character propoerties is assigned to each text-line and iii) the *classification* module for distinguishing the printed from the handwritten text-lines. An overview of the system is shown in figure 1.

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

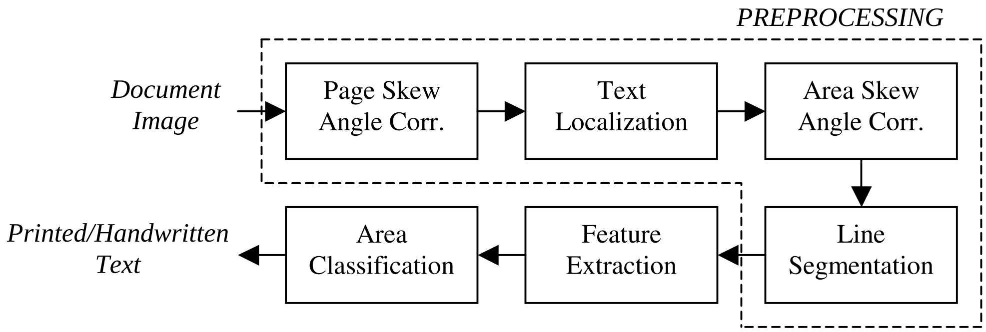

Figure 1. System layout

2.1 Preprocessing

The preprocessing stage aims at the localization and segmentation of the text areas of the document. Existing algorithms [8-10] are applied in order to extract a set of text-lines that compose the document. In this approach, it is assumed that there are no images, graphics or banners in the document.

An important preproocessing factor is the *skew angle correction*. The skew angle estimation is performed by employing its horizontal histogram and the Wigner-Ville Distribution (WVD), an approach described in detail in [8]. Specifically, the maximum intensity of the WVD of the horizontal histogram of a document image is used as the criterion for its skew angle estimation. At first, the skew angle correction algorithm is applied *on the page-level* providing a rough estimation for the whole page. Then, the algorithm is applied *on the text-area-level*, for fine tuning the estimation for each area. This two-step approach is necessary for two reasons: 1) in many cases the handwritten text can be of different orientation in comparison to the printed text (e.g., handwritten notes on a printed page), 2) the orientation of handwritten text may be variable within the same page.

The *discrimination and localization of text areas* is performed based on the algorithm described in [9]. Specifically, a stage of segmentation is performed where the constrained run-length algorithm (CRLA) [11], also known as 'smearing', is used. The document is segmented in smaller areas, called first-order connected components (CC). Before going further, the first-order CCs that satisfy any one of the following criteria are eliminated [12]:

(a) The area of their corresponding Bounding Boxes (BB) is lower than the value Amin=100 pixels. Those CCs are assumed to be noise.

(b) Their aspect ratio, i.e. the ratio between the width and the height of the corresponding BB, is lower than 1.0/20.0. This region, most probably, does not contain text information, e.g., a vertical line.

(c) The aspect ratio is greater than 20.0/1.0. It may be, e.g., a horizontal line.

In this study we assume that the document image contains no images but it may include vertical and horizontal strokes. Since those strokes are already limited (from the previous procedure), we expect that the remaining areas will be blocks of the same type of text, which proved to be true in our experiments.

For the *line segmentation* task, a very simple algorithm [10] was used, which is a variation of a well-known approach [13]. This variation is employed since it combines ease of implementation and high accuracy results. Summarizing, The preprocessing stage provides a set of text-lines either printed or handwritten that compose the document. Some of the text lines may contain just one word or a few words.

2.2 Feature extraction

The main idea of our approach is to take advantage of the structural properties that help humans discriminate printed from handwritten text. In more detail, the height of the printed characters is more or less stable within a text-line. On the other hand, the distribution of the height of handwritten characters is quite diverse. These remarks stand also for the height of the main body of the characters as well as the height of both ascenders and descenders. Thus, the ratio of ascenders' height to main body's height and the ratio of descenders' height to main body's height is more likely to be stable in printed text and variable in handwriting.

The extraction of the feature vector of each text-line, is based on the upper–lower profile (i.e., the position of both the first and last black pixels on each column), which essentially provides an outline of the text-line. Consider that the value of the element in the *m*-th row and *n*-th column of the text line matrix is given by a function *f*:

$$f(m,n) = a_{mn}$$

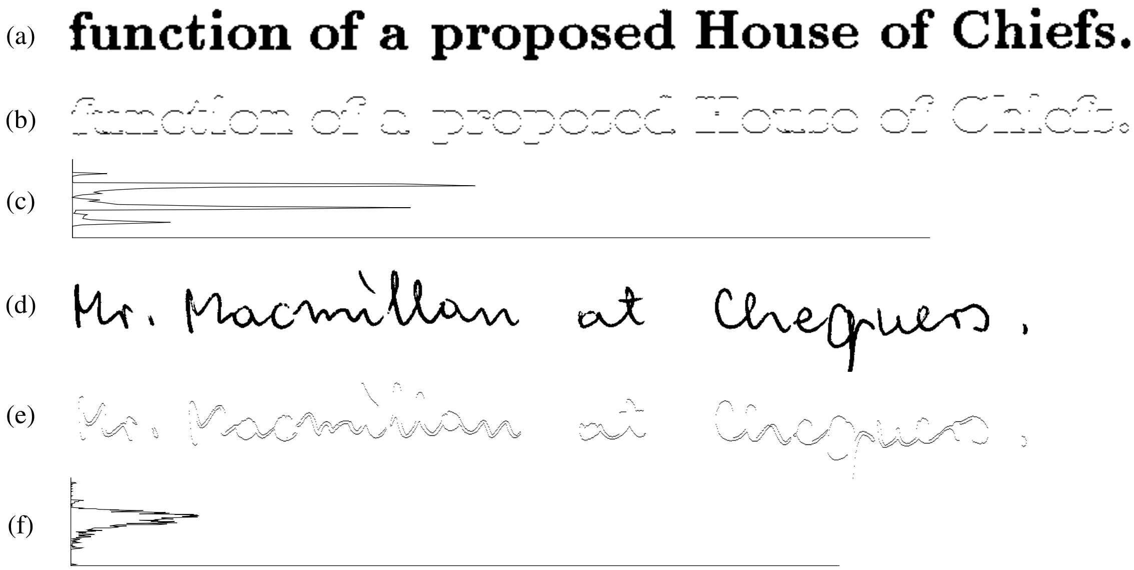

Figure 2. Examples of upper-lower profile: (a) a printed text-line, (b) its upper-lower profile, (c) the horizontal histogram of the profile, (d) a handwritten text-line, (e) its upper-lower profile, (f) the horizontal histogram of the profile,

where α_{mn} takes binary values (i.e., 0 for white pixels and 1 for black pixels). The upper-lower profile P of an image is:

$$P(x) = \left\{ \begin{array}{l} (J1, J2) : \sum_{i=0}^{J1-1} f(i,x) \equiv 0 \;\&\; \sum_{i=J2+1}^{height} f(i,x) \equiv 0 \;\& \\ \& \; f(J1,x) = f(J2,x) \equiv 1 \end{array} \right\},$$

$$x \in [0, length\_of\_image)$$

Using the horizontal histogram of the upper-lower profile, we are able to estimate the height of (i) the main body zone, (ii) the ascender zone, and (iii) the descender zone. In particular, the peak of the horizontal histogram of the upper-lower profile located above the middle of the profile (upper peak) and the corresponding peak below the middle of the profile (lower peak) define the main body zone. The ascender zone is defined above the upper peak and the descender zone is defined below the lower peak.

Figure 2 shows examples of upper-lower profiles for both printed and handwritten text-lines. As can be seen, the detection of the main body, ascender, and descender zones is much more obvious using the horizontal histogram in the case of machine-printed text.

The features used to characterize each text-line are: (i) the ratio of ascender zone to main body zone, (ii) the ratio of the descender zone to the main body zone, and (iii) the ratio of the area to the maximum value of the horizontal histogram of the upper-lower profile.

2.3 Classification method

The classification method used in the following experiments is discriminant analysis, a standard technique of multivariate statistics. The mathematical objective of this method is to weight and linearly combine the input variables in such a way so that the classes are as statistically distinct as possible [14]. A set of linear functions (equal to the input variables and ordered according to their importance) is extracted on the basis of maximizing between-class variance while minimizing within-class variance using a training set. Then, class membership of unseen cases can be predicted according to the *Mahalonobis* distance from the classes' centroids (the points that represent the means of all the training examples of each class). The Mahalanobis distance d of a vector x from a mean vector m is as follows:

$$d^2 = (x-m)' C_x^{-1} (x-m)$$

where C_x is the covariance matrix of x. This classification method also supports the calculation of posterior probabilities (the probability that an unseen case belongs to a particular group) which are proportional to the Mahalanobis distance from the classes centroids. In a recent study [15], discriminant analysis is compared with many classification methods (coming from statistics, decision trees, and neural networks). The results reveal that discriminant analysis is one of the best compromises taking into account the classification accuracy and the training time cost. This old and simple statistical algorithm performs better than many modern versions of

Table 1. ANOVA tests for the proposed features ($p<0.0001$)

Feature	r^2**(%)**
Ascender zone / Main body zone	91.3
Descender zone / Main body zone	93.2
Area / Peak value	98.0

statistical algorithms in a variety of problems. Given that it is an easy-to-implement method, it provides an ideal classification algorithm for testing new feature sets.

3. Evaluation

The proposed approach has been tested on document images taken from two databases: IAM-DB (English text) and GRUHD (Greek text). Both databases contain mixed documents (machine-printed and handwritten text areas). 50 document images were randomly selected and preprocessed (see Section 2.1) resulting a series of text-lines. For each text-line a vector with the proposed features was calculated. Then, 10-fold cross-validation was applied. The text-lines were divided into ten non-overlapping sets. Each time a classification model was calculated with training examples taken from one set and evaluated on the remaining sets. This procedure was repeated ten times, each time using a different set as training examples. The average classification accuracy was 98.2 %. A great part of errors come from handwritten text-lines of short length (usually just one word) erroneously classified as printed text.

Another important point is that the proposed approach requires minimal training sets in order to achieve very high accuracy. Using just two training examples for each class (i.e., two text-lines for machine-printed and two text-lines of handwritten text as training set) accuracy of 97.9% was achieved.

The significance of the proposed features was tested using the statistical method analysis of variance (aka ANOVA). Specifically, ANOVA tests whether there are significant differences among the classes with respect to the measured values of a particular feature. Table 1 shows the results of this analysis for each feature. r^2 measures the percentage of the variance among feature values that can be predicted knowing the class of the text-line. So, the greater the r^2 value, the most significant the feature. As can be seen, the area to peak value ratio of the horizontal histogram of the upper–lower profile proves to be the most reliable feature.

4. Conclusions and future work

A text identification system was presented, able to discriminate between machine-printed and handwritten text-lines. The proposed solution can handle document pages, identifying text areas and splitting each area into text-lines. A set of simple and easy-to-compute structural characteristics is introduced. According to the presented experiments, the proposed features capture significant amount of the differences between machine-printed and handwritten text providing a good solution for this task.

Preliminary experiments on document images taken from two databases of latin-style languages prove that remarkable results can be aquired using minimal training examples from each class. On the other hand, handwritten text-lines of short length prove to be the most difficult case.

As future work directions, we plan to compare the presented approach with methods that are based on different text-areas (e.g., word-level) and to further investigate the properties of the structural characteristics that are optimal for the discrimination between machine-printed and handwritten text. Moreover, an integrated system should be able to deal with document images that include any type of graphics/fonts.

5. References

[1] Y.Zheng, H.Li, D.Doermann, "Text identification in Noisy Document Images Using Markov Random Field", *Proc of 7th ICDAR,* 2003, pp.599-603.

[2] K.C.Fan, L.S.Wang and Y.T.Tu, "Classification of machine-printed and handwritten texts using character block layout variance", *Pattern Recognition,* 31(9), 1998, pp.1275-1284.

[3] V.Pal and B.B.Chaudhuri, "Machine-printed and handwritten text lines identification", *Pattern Recognition Letters,* 22, 2001, pp.431-441.

[4] K.Nitz, W.Cruz, H.Aradhye, T.Shaham and G.Myers, "An Image-based Mail Facing and Orientation System for Enhanced Postal Automation", *Proc of 7th ICDAR,* 2003, pp.694-698.

[5] H.Ma and D.Doermann, "Gabor Filter Based Multi-class Classifier for Scanned Document Imagrs", *Proc of 7th ICDAR,* 2003, pp.968-972.

[6] U. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition". *Proc. 5th Int. Conference on Document Analysis and Recognition, ICDAR'99*, 1999, pp. 705 – 708.

[7] E.Kavallieratou, N.Liolios, E.Koutsogeorgos, N.Fakotakis, G.Kokkinakis, ``The GRUHD database of Modern Greek Unconstrained Handwriting", *In Proc. ICDAR* ,2001 v.1, 2001, pp.561-565.

[8] E. Kavallieratou, N. Dromazou, N. Fakotakis and G. Kokkinakis, "An Integrated System for Handwritten Document Image Processing", *IJPRAI, International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 17, No. 4, 2003, pp. 101-120.

[9] E.Kavallieratou, D.C.Balcan, M.F.Popa, and N.Fakotakis, "Handwritten Text Localization in Skewed Documents", *ICIP'2001*, 2001, pp.1102-1105.

[10] E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, "Un Off-line Unconstrained Handwritting Recognition System", *International Journal of Document Analysis and Recognition*, no 4, 2002, pp. 226-242.

[11] Wahl, F. M., Wong, K. Y. and Casey, R. G.: "Block segmentation and text extraction in mixed text/image documents", *Comput. Graph. Image Processing,* 20, pp. 375-390, 1982.

[12] L. A. Fletcher and R. Kasturi "A robust algorithm for text string separation from mixed text/graphics images", *IEEE Trans, PAMI*-10, (6), pp. 910-918, 1988.

[13] M.Shridar and F.Kimura, "Handwritten address interpretation using word recognition with and without lexicon", *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, v 3, 1995, pp.2341-2346.

[14] Eisenbeis, R. and R. Avery, *Discriminant Analysis and Classification Procedures: Theory and Applications*, Mass.: D.C. Health and Co., Lexington, 1972.

[15] Lim, T., W. Loh, and Y. Shih, "A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Algorithms", *Machine Learning* 40 (3). 2000, pp. 203-228.

0-7695-2187-8/04 $20.00 © 2004 **IEEE**
IEEE COMPUTER SOCIETY