Automatic discrimination between printed and handwritten text in documents

Lincoln Faria da Silva, Aura Conci
Instituto de Computação
Universidade Federal Fluminense - UFF
Niterói, Brasil
{lsilva, aconci}@ic.uff.br

Angel Sanchez
Departamento de Ciencias de la Computacion
Universidad Rey Juan Carlos,
Madrid, Spain
angel.sanchez@urjc.es

***Abstract*—Recognition techniques for printed and handwritten text in scanned documents are significantly different. In this paper we address the problem of identifying each type. We can list at least four steps: digitalization, preprocessing, feature extraction and decision or classification. A new aspect of our approach is the use of data mining techniques on the decision step. A new set of features extracted of each word is proposed as well. Classification rules are mining and used to discern printed text from handwritten. The proposed system was tested in two public image databases. All possible measures of efficiency were computed achieving on every occasion quantities above 80%.**

***Keywords*-Data Mining; document analysis; text identification; optical characters recognition; Machine Vision**

I. Introduction

Great number of applications use documents presenting printed text and handwriting. Old documents, petitions, requests, applications for college admission, letters, requirements, memorandums, envelopes and bank checks are some examples. A considerable obstacle to optical character recognition (OCR) systems is the mixture of printed and handwritten text in the same image. Each text type should be processed using different methods in order to optimize the recognition accuracy.

Previous works addressed the problem of identifying each type by various classification techniques. These works utilize neural networks [1-7], employ linear polynomial for discrimination function [8], Fisher [9-12] and tree classifiers [13-14], Hidden Markov Model (HMM) [15] or minimal distance classifiers [16-17]. In this paper we propose the use of classification rules mining by the WEKA tool [23]. This enables us to visualize best rules from a group of possible classifiers using features extracted from each word of the document. The main advantage of this, compared with other classifiers, is its accuracy, efficiency, simplicity and the low computation complexity. When the classification is performed by words and not by line, it is possible to analyse more complex pages which mix in the same line both type of characters. However, all documents to be classified are supposed to be aligned with the scanner. The implemented system is concerned with documents presenting adequate orientation on the acquisition step, not to skewed one.

Document image is firstly preprocessed by various techniques. Then the text is segmented at word level when each word is surrounded by a bounding box (BB). Afterward features are extracted from these BBs. The classification rules decide whether a BB contains printed or handwritten text. Two public image databases are used to verify the implemented system. Both present very satisfactory results permitting evaluation of its robustness.

This paper is organized as follows: In section 2 each step of the overall system is presented. Section 3 considers the training, tests and results. Finally, section 4 summarizes the conclusions and future improvements.

II. The proposed approach

This section presents the proposed system. It describes the type of document processed, the applied image processing techniques, the segmentation of the text in words, the extracted features from these words, and the classification process executed by the system. Figure 1 shows an overview of the system. It has four main steps: preprocessing, text segmentation in words, feature extraction and classification.

A. Document types

The developed system considers application forms for various objectives, such as subscription forms, research questionnaires or preprinted memorandums. Blank regions, lines, printed and handwritten words can be found all over these documents. However, they do not present logos, figures, tables, graphs or another type of element. Figure 2 shows an example of possible images to be processed. Note that for systems performing classification at line level it is not possible the combination of written and printed in the same line of the documents.

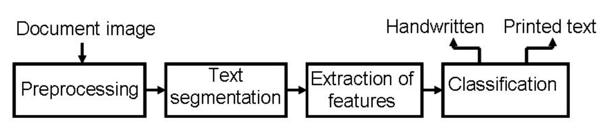


Fig. 1. Overview of the system

1530-1834/09 $25.00 © 2009 IEEE
DOI 10.1109/SIBGRAPI.2009.40

B. Preprocessing steps

The preprocessing phase prepares the acquired image to segmentation in words. Four operations are accomplished in this step. Figure 3 illustrates the steps of this phase.

Firstly, a 3x3 median filter is applied to decrease the noise in the image. This noise can appear on document acquisition or digitalization. Figure 4 shows a region of a document before (a) and after (b) this operation.

Secondly, the text is separated from background by automatic thresholding. The Otsu bi-level approach [24] is used to define the threshold. Figure 5 shows this operation.

Then the horizontal lines used as a guide in the writing process (Figure 2) must be removed. In this case an simple line extraction algorithm considering the connected elements of the lines is applied [27]. Figure 6 shows this operation.

Finally, the document image is submitted to morphologic opening (that is an erosion followed by dilation) by a 3x1 symmetrical structuring element. This is designed with two purposes: eliminating reminiscent noises of the previous phase and soften vertical contours of the text. These are essential for proper computation of some features to be extracted. Figure 7 shows this.

C. Word segmentation step

The input of this step is a binary document and the output is the localization of each word by its coordinates and limits (BB). Adequate and correct text segmentation in word is the most important step for intelligent OCR and for text identification as handwritten or printed, as well.

In order to localize each word, the system performs the extraction of connected components. These are the first candidate of BB and are identified by boxes as shown in Figure 8. Box in the same text line and with distance less than half of the average width of boxes or with overlapping pixels are united forming words (see Figures 9 and 10). Average width is calculated by (1):

$$Distance = \frac{\sum_{i=0}^{k} L_i}{2k} \tag{1}$$

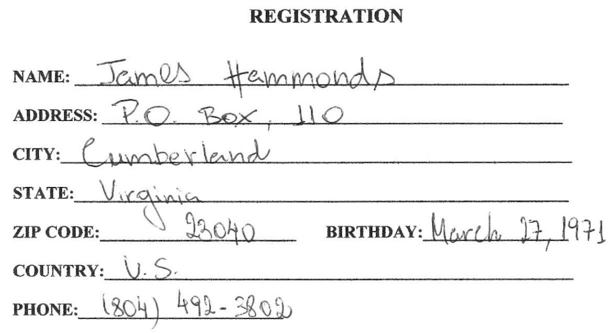
where k is the number of boxes and L_i the lengths of all boxes in the image.

D. Extracted features

Eleven features are defined and extracted to each BB individually.

1) Deviation of the Widths (D_W), Heights (D_H) and Areas(D_A): Initially, the averages of the widths (M_W), of the heights (M_H) and of the areas (M_A) of the bounding boxes, on document image, are calculated. For each BB, the module of difference of its width (W) by average of the widths (M_W) is stored as a feature. The same is made with its heights (H) and its area (A). Equations (2-4) represent these new features that we proposed in our work.

$$D_W = |W - M_W| \tag{2}$$


REGISTRATION

NAME: James Hammonds

ADDRESS: P.O. Box, 110

CITY: Cumberland

STATE: Virginia

ZIP CODE: 23040 BIRTHDAY: March 27, 1971

COUNTRY: U.S.

PHONE: (804) 492-3802

Fig. 2. A sample of a used database

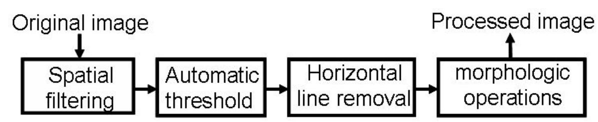


Fig. 3. Preprocessing steps

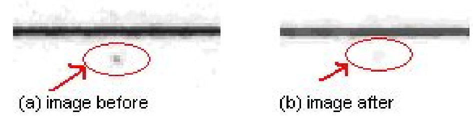


Fig. 4. Application of the median filter

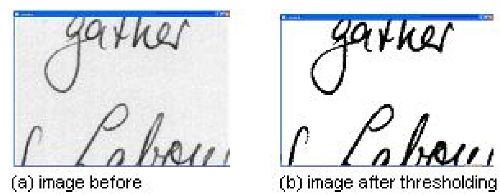


Fig. 5. Automatic threshold of the image

Fig. 6. Horizontal lines removal

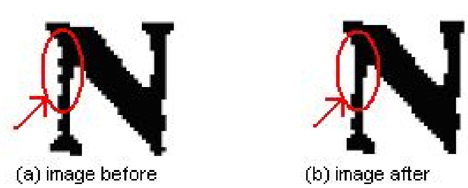


Fig. 7. Smoothed vertical contours

$$D_H = |H - M_H| \tag{3}$$

$$D_A = |A - M_A| \tag{4}$$

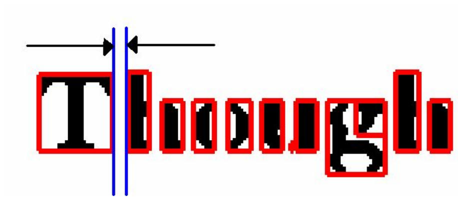


Fig. 8. Connected components surrounded by box candidates to words

Fig. 9. BB from boxes united forming words

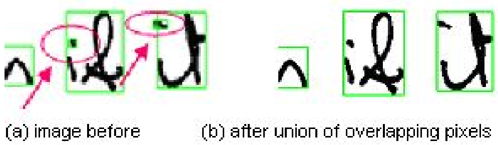


Fig. 10. Union of overlapping box to form BBs

2) Density: We call "density" the black pixels area inside the BB. This feature is computed by the relation between area of black pixels inside the BB and the total BB area. This can be represented by (5):

$$Density = \frac{of\ BB\ black\ pixels}{BB\ area} \tag{5}$$

3) Vertical Projection Variance: For computation of this feature, the vertical projection of black pixels inside the BB is evaluated (Figure 11). Then, the variance of the vertical coordinates of the profile of this vertical projection is calculated as a measure of homogeneity of the projection profile. Figure 12 shows the vertical projection profile of Figure 11.

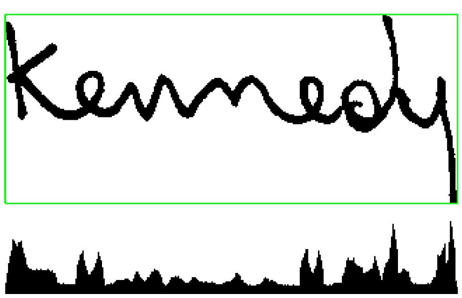


Fig. 11. Vertical projection of an BB

4) Major Horizontal Projection Difference: For computing this feature, the horizontal projection of the black pixels inside the BB is executed. Figure 13 shows this projection for a handwritten word. Then the major absolute difference of the abscissas of adjacent pixels of this profile is computed. Figure 14 shows the projection profile of the BB in Figure 13.

Fig. 12. BB vertical projection profile

Fig. 13. Horizontal projection of an BB

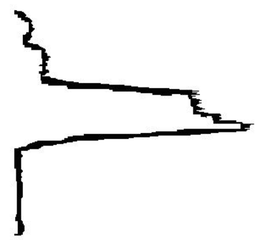


Fig. 14. BB horizontal projection profile

5) Pixels Distribution: In order to analyze its pixels distribution, the BB is divided in two by a horizontal line as indicated by the red line on Figure 15. Then the BB height is decreased by 10 pixels as shown by the green lines on the Figure 15. Then the density of the upper part (UD) and of the lower part (LD) is calculated by (6). Finally, the module of the difference between these densities is stored as a feature proposed in our work.

$$Pixels\ Distribuition = |UD - LD| \tag{6}$$

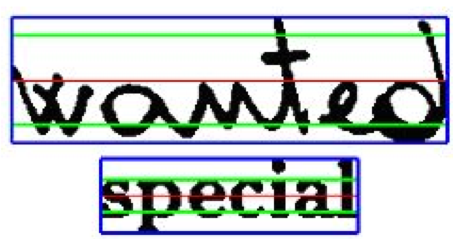


Fig. 15. BBs horizontal division

6) Bottom Line: The number of pixels of the word inside the BB that intercept its inferior border line is computed. Figures 16 and 17 shows this for a handwritten word and a printed word, respectively.

The relation between the total number of pixels in such condition and the width of the BB is stored as a new feature which we call "Bottom Line".

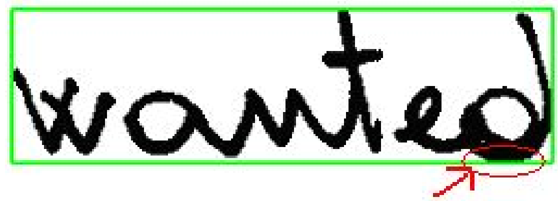


Fig. 16. Bottom line of pixels of a handwritten

Fig. 17. Bottom line of pixels of a printed word

7) Black Pixels of Each Line: Each line of the BB has the number of black pixels computed and divided by BB width. The sum of the results of these divisions is stored as a feature proposed for us.

8) Vertical Edges: The vertical edges of the characters compounding each word are detected for the spatial filter masks presented in Figure 18. These masks are based on the directional Prewitt filter [27]. Figure 19 shows the result of application of these masks in a handwritten and in a printed word. The sum of the lengths of all the detected vertical character contours inside the BB divided by the BB area is another new proposed feature.

1	0	-1
1	0	-1
1	0	-1
1	0	-1
1	0	-1
1	0	-1
1	0	-1

-1	0	1
-1	0	1
-1	0	1
-1	0	1
-1	0	1
-1	0	1
-1	0	1

Fig. 18. Spatial filter masks

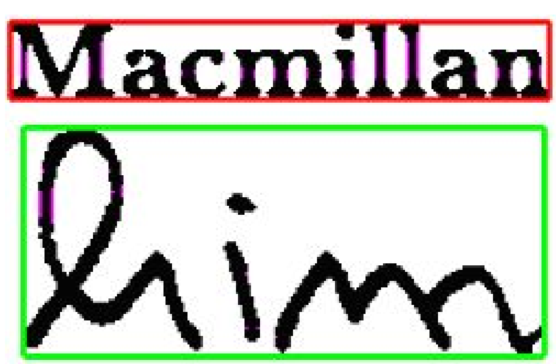


Fig. 19. Vertical edges of a handwritten and printed word

9) Major Vertical Edge: The vertical edges inside the BB obtained of last feature is used here. However, only the bigger vertical edge (that is, the one with greater number of pixel) is considered to compute this feature. Figure 20 shows this. Thus the relation between the number of pixels of the major vertical edge inside the BB and its height is a new feature, which we call "Major Vertical Edge".

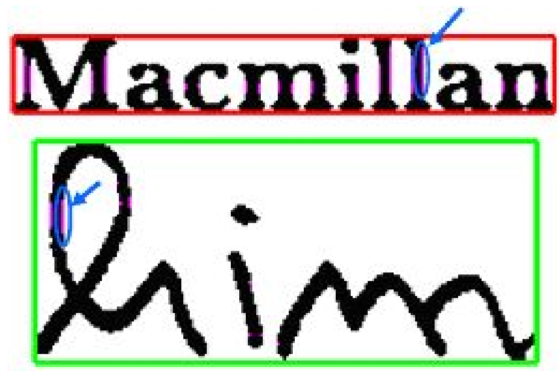


Fig. 20. Major vertical edge of words

E. System classification

For each document, the features described in the last section are calculated for each BB and the classification rules, mined in the training phase, are used to decide if a BB contains printed or handwritten word.

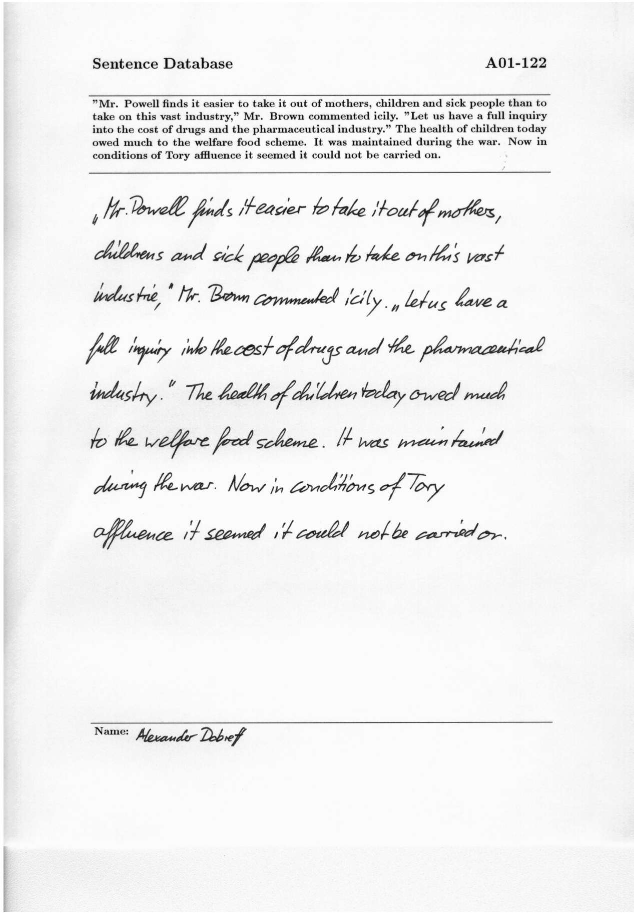
III. Experiments

Tests have been performed by using two public databases. One of these databases is the IAM Database version 3.0 [18-21]. This database is formed for 1539 documents containing print and handwritten text. An document of this database can be seen in Figure 21. Regions of printed and handwritten words of this database is easily separable. Only in the last line of the document appears both types of elements to be classified in the same line. Moreover, this database presets no auxiliary lines to fill or to supply with written texts. This characteristic facilitates the identification and classification of each types of words. However, this is not so useful to show all the possibility of the implemented system.

Therefore a new database was constructed in [27] and avaiable in [28]. It is formed of 121 application form images filled by 121 voluntarily with various education level, age and social status. Each filled form was scanned in 300 dpi (256 gray level) and converted to BMP format. Differently of the IAM Database, this is a challenge. An example of parts of this database can be seen in Figures 2 and 6. It presents printed and handwritten words in the same line and many horizontal lines to be used as a guide in the manual process of filling in.

A. Training stage

The classification rules (Table I) are obtained in the training phase. To archive this, the features and the correct classification (training set) of some elements of a database are used as


Sentence Database A01-122

"Mr. Powell finds it easier to take it out of mothers, children and sick people than to take on this vast industry," Mr. Brown commented icily. "Let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

,,Mr. Powell finds it easier to take it out of mothers, childrens and sick people than to take on this vast industrie," Mr. Brown commented icily. ,,let us have a full inquiry into the cost of drugs and the pharmaceutical industry." The health of children today owed much to the welfare food scheme. It was maintained during the war. Now in conditions of Tory affluence it seemed it could not be carried on.

Name: Alexander Dobref

Fig. 21. IAM Database form

input file of the WEKA tool. This is a free machine learning collection of algorithms used for mining [25-26] classification rules. The set of more significant features is changed according to the used training set. This runs the CfsSubsetEval valuator with the GeneticSearch method in the WEKA tool. Table II shows the features found as the most important ones.

TABLE I
SOME CLASSIFICATION RULES OF THE IAM DATABASE

Antecedent	Class
IF{(**Major Vertical Edge** $\leq$ 0.422) and (**Vertical Projection Variance** $\leq$ 99)}	THEN {**handwritten**}
IF{(**Major Vertical Edge** $>$ 0.422) and (**Major Horizontal Projection Difference** $>$ 27) and (**Pixels Distribution** $\leq$ 594)}	THEN {**printed**}
IF{(**Major Horizontal Projection Difference** $\leq$ 20)	THEN {**handwritten**}

B. Tests and results

For the realization of the test using the IAM Database, 20 images were chosen randomly and separated in 10 subsets to be used by the k-fold cross validation method (i.e. k = 10).

TABLE II
MOST INSIGNIFICANT FEATURES

Density
Vertical Projection Variance
Major Horizontal Projection Difference
Pixels Distribution
Vertical Edges
Major Vertical Edge

Thus each subset was formed for two of these images (forms). False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN) were calculated to allow adequate quantitative comparison with others works by using quality evaluator parameter. Table III shows the result of the test using the IAM Database 3.0. All the BBs were correctly classified (i.e. accuracy and precision equal 100%) in 45% of the images of this database.

For the testes using the new developed database, 24 images were randomly chosen and separated in 3 subsets required by the k-fold cross validation method, when k = 3. Thus each subset was formed by eight of these images (forms). Table IV shows the result of the test using our database [28]. accuracy and precision equal 100% were achieved by the system in 33% of the images in such database.

In Tables III and IV, the lines "Average of the accuracies/precisions" of the printed/handwritten words contain the average of the accuracies/precisions of the ten subsets. The standard deviation of the accuracies/precisions is also calculated in relation to accuracies and precisions of the ten subsets.

TABLE III
TEST RESULT IN THE IAM DATABASE 3.0

Quality evaluator parameter	Printed	Hand-written
Total number of words	1404	2029
Accuracy	97.51%	97.54%
Precision	96.48%	98.26%
False Positive	50	35
False Negative	35	50
True Positive	1369	1979
True Negative	1979	1369
Sensibility	0.97	0.97
Specificity	0.97	0.97
Average of the accuracies	97.55%	98.09%
Average of the precisions	96.70%	98.10%
Accuracies(standard deviation)	1.61	2.03
Precisions (standard deviation)	4.41	1.38
Minimum accuracy	91.18%	91.01%
Minimum precision	81.82%	93.85%
Range of the accuracies	8.82%	8.99%
Range of the precisions	18.18%	6.15%

Accuracy, Precision, Sensibility and Specificity are calculated by (7-10):

$$Accuracy = \frac{CC}{\#\ of\ BBs\ of\ the\ c_k\ class} \quad (7)$$

$$Precision = \frac{CC}{\#\ of\ BBs\ classified\ as\ of\ the\ c_k\ class} \quad (8)$$

$$Sensibility = \frac{TP}{TP + FN} \quad (9)$$

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

where CC is the number of correctly classified BBs of the c_k class.

A computer equipped with the processor AMD Athlon MP 900Mhz was used for the tests and the system carried out, approximately, 184.04 BI (billions of instructions) for each processed image. BI is an evaluation concept proposed for efficient time comparison [22]. The value of 184.04 is obtained from the following calculation: (74*2487)/1000, where 74 is the time average, in seconds, consumed for processing of one image. 2487 is the quantity of instructions, in millions by second, which the processor is capable of executing, called of MIPS (millions of instructions by second), obtained after application of a processor arithmetic test.

TABLE IV
TEST RESULT IN OUR DATABASE [28]

Quality evaluator parameter	Printed	Hand-written
Total number of words	600	1329
Accuracy	97.17%	99.47%
Precision	98.81%	98.73%
False Positive	7	17
False Negative	17	7
True Positive	583	1322
True Negative	1322	583
Sensibility	0.97	0.99
Specificity	0.99	0.97
Average of the accuracies	97.17%	99.46%
Average of the precisions	98.85%	98.75%
Accuracies(standard deviation)	1.89	0.76
Precisions (standard deviation)	1.58	0.80
Minimum accuracy	88.00%	96.43%
Minimum precision	92.59%	95.35%
Range of the accuracies	12.00%	3.57%
Range of the precisions	7.41%	4.65%

C. Comparisons

Results of the here proposed methodology and others allows to observe many advantages of our approach. Whereas our system classifies at word level, Kavallieratou et al. [16-17] classify at line level. These works used the same IAM database (that is the database used by Kavallieratou and Stamatatos [16]). This is the unique database that is really available. For this reason this is the only one used here for comparison. However, their technique does not permit to distinguish printed from handwritten in same text line. Moreover, a simple comparison concerning the accuracy is not possible because they just announce the number 97.9% without specific description of what measure it is related. As our results on Table 3 shows, our average accuracy is almost the same, exactly 97.55% for printed word and 98.09% for handwritten words. In addition, for almost half (exactly 45%) of the IAM Database, an accuracy of 100% was achieved by our methodology on discrimination between printed and handwritten words. Table V summarizes such comparison.

TABLE V
COMPARISON WITH KAVALLIERATOU ET AL.[16, 17]

	Kavallieratou et al.	Our methodology
Database used	IAM-DB 3.0	IAM-DB 3.0
Classification	Lines level	Words level
Validation method	k-fold cross v. with k = 10	k-fold cross v. with k = 10
Accuracy	97.9%	100%

IV. CONCLUSIONS

In this paper we proposed a set of new features to be extracted from images and an approach to find classification rules for discrimination printed and handwritten text in documents. The system was implemented and tested. A new database was completely classified and published [28].

The developed system was tested in two public image databases [18,28], applying the k-fold cross validation method. Experiments show that our methodology is robust and applicable to a majority of document types. The database created here [28] is an important contribution to development of future comparisons. Moreover, the extracted features to represent the printed and handwritten words proposed make the system independent of the document layout in the discrimination task. It is possible to analyze texts with printed and handwritten words in the same line because the classification is performed by words and not only by line as previous works. Finally, as our approach presents very good results by using only classification rules it is also less time consuming then all others methodologies used until now.

In the future we plane to implement other classification techniques (such as minimum distance, Support Vector Machines, neural networks, and those based in on fuzzy logic) for comparisons with the proposed classification techniques. Moreover, new features will be extracted of the BBs exploring the regularity of the printed words and the absence of this on the handwritten words. Other future improvement could be a hybrid text segmentation, part top down and part bottom up. The part top down segments the text in lines, and the part bottom up segments the segmented text lines in words. Finally, we consider to improve this system for performing the discrimination in a larger variety of document layouts including other text orientations on the page (not only horizontal), documents with tables, figures, graphs and other elements as annotations among the printed words could be considered as well.

Acknowledgment

We acknowledge the grants provided by Brazilians agencies CAPES and CNPq.

References

[1] S. Imade, S. Tatsuta, and T. Wada, *Segmentation and Classification for Mixed Text/Image Documents Using Neural Network* , Proceedings of the Second International Conference on Document Analysis and Recognition, 20-22 Oct., pp. 930 - 934, 1993.

[2] S. Violante, R. Smith, and M. Reiss, *A Computationally Efficient Technique for Discriminating Between Hand-Written and Printed Text*, IEEE Colloquium on Document Image Processing and Multimedia Environments, 2 Nov.,pp. 17/1 - 17/7, 1995.

[3] K. Kuhnke, L. Simoncini, and Z. M. Kovacs-V, *A System for Machine-Written and Hand-Written Character Distinction*, Proceedings of the Third International Conference on Document Analysis and Recognition,v. 2,14 - 16 Aug.,pp 811 - 814, 1995.

[4] J. E. B. Santos, B. Dubuisson, and F. Bortolozzi, *A Non Contextual Approach for Textual Element Identification on Bank Cheque Images*, IEEE International Conference on Systems, Man and Cybernetics, v. 4, pp 6 - 9, 2002.

[5] J. E. B. Santos, B. Dubuisson, and F. Bortolozzi, *Characterizing and Distinguishing Text in Bank Cheque Images*, Proceedings XV SIBGRAPI, pp. 203 - 209, 2002.

[6] F. Farooq, K. Sridharan, and V. Govindaraju, *Identifying Handwritten Text in Mixed Documents*, ICPR 2006, 18th International Conference on Pattern Recognition, v. 2, pp. 1142 - 1145, 2006.

[7] J. Koyama, M. Kato, and A. Hirose, *Local-spectrum-based distinction between handwritten and machine-printed characters*, 15th IEEE International Conference on Image Processing, 12-15 Oct., pp. 1021 - 1024, 2008.

[8] J. Franke, and M. Oberlander, *Writing Style Detection by Statistical Combination of Classifiers in Form Reader Applications*, Proceedings of the 2nd Intern. Conference on Document Analysis and Recognition, pp. 581 - 584, 1993.

[9] S. N. Srihari, Y. C. Shin, V. Ramanaprasad, and D. S. Lee, *A System to Read Names and Addresses on Tax Forms*, Proceedings of the IEEE, v. 84, n 7, pp. 1038 - 1049. DOI: 10.1109/5.503302, 1996.

[10] Y. Zheng, H. Li, and D. Doermann, *The Segmentation and Identification of Handwriting in Noisy Document Images*, , Document Analysis Systems V, Lecture Notes in Computer Science, v. 2423, pp. 95-105, 2002.

[11] Y. Zheng, H. Li, and D. Doermann, *Text Identification in Noisy Document Images Using Markov Random Field*, Proceedings of the Seventh International Conference on Document Analysis and Recognition, v. 1, pp. 599 - 603, 2003.

[12] Y. Zheng, H. Li, and D. Doermann, *Machine Printed Text and Handwriting Identification in Noisy Document Images*, , IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 26, n 3, pp. 337 - 353, 2004.

[13] U. Pal, and B. B. Chaudhuri, *Automatic separation of machine-printed and hand-written text lines*, ICDAR '99. Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 645-648, 1999.

[14] U. Pal, and B. B. Chaudhuri, *Machine-printed and Hand-written Text Line Identification*, Pattern Recognition Letters, v. 22, n 3 - 4, pp. 431 - 441, 2001.

[15] J. K. Guo, and M. Y. Ma, *Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models*, , Proceedings. Sixth International Conference on Document Analysis and Recognition, pp. 439 - 443, 2001.

[16] E. Kavallieratou, and S. Stamatatos, *Discrimination of Machine-Printed from Handwritten Text Using Simple Structural Characteristics*, Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, v. 1, 23 - 26 Aug., pp.437 - 440, 2004.

[17] E. Kavallieratou, S. Stamatatos, and H. Antonopoulou, *Machine-Printed from Handwritten Text Discrimination*, IWFHR-9 2004, 9th Intern. Workshop on Frontiers in Handwriting Recognition, 26-29 Oct., pp. 312 - 316, 2004.

[18] U. Marti, and H. Bunke, *The IAM-database: an English Sentence Database for Off-line Handwriting Recognition*, Int. Journal on Document Analysis and Recognition, v. 5, n1, pp. 39-46, 2002.

[19] U. Marti, and H. Bunke, *A full English sentence database for off-line handwriting recognition*, ICDAR '99, Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 705-708, 1999.

[20] U. Marti, and H. Bunke, *Handwritten Sentence Recognition*, Proceedings. 15th International Conference on Pattern Recognition, v. 3, pp. 463-466, 2000.

[21] M. Zimmermann, and H. Bunke, *Automatic Segmentation of the IAM Off-line Database for Handwritten English Text*, Proceedings of the 16th International Conference on Pattern Recognition, v. 4, pp. 35 - 39, 2002.

[22] A. Brazil, *Path Relinking and AES Cryptography in Color Image Steganography*, IC/UFF- M. Sc Thesis, 2008. available in: http://www.ic.uff.br/PosGraduacao/Dissertacoes/375.pdf

[23] *WEKA, 1999*, documentation available in: http://www.cs.waikato.ac.nz/ml/weka/

[24] N. Otsu, *A Threshold Selection Method from Gray-Level Histograms*, IEEE Transactions on Systems, Man and Cybernetics, v. 9, n 1, pp. 62 - 66, 1979.

[25] J. Han, and M. Kamber, *Data mining: Concept and Techniques*, ed. Morgan Kaufmann, 2001.

[26] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, AI Magazine, v. 17, n 3, 1996, pp. 37-54.

[27] L. F. Silva, *Distinção o Automática de Texto Impresso e Manuscrito em uma Imagem de Documento*, IC/UFF- M. Sc Thesis, 2009. available in: http://www.ic.uff.br/PosGraduacao/Dissertacoes/411.pdf

[28] Database available in: http://visual.ic.uff.br/analisededocumentos/pt/bancoimagens.htm