Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 10 (2013) 831 – 839

Procedia Technology

International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

Handwritten and Printed Word Identification Using Gray-Scale Feature Vector and Decision Tree Classifier

Samir Malakar[a]*, Rahul Kumar Das[a], Ram Sarkar[b], Subhadip Basu[b], Mita Nasipuri[b]

[a]*Deparment of Master of Computer Application, MCKV Institute of Engineering, Howrah, India*
[b]*Department of Computer Science and Engineering, Jadavpur University, Jadavpur, Kolkata, India*

Abstract

Document image analysis is one of the important steps towards a paper free world. An effective Optical Character Recognition (OCR) system would be helpful for achieving this fit. But the next question may arise that whether a single OCR system will be sufficient for encoding both handwritten and printed text or not. So to come out of this dilemma, the work as reported here determines the category of a word from the document images containing words both in handwritten and printed forms. A 6-elements feature set is estimated from each gray level image and then these features are ranked based on discriminatory capabilities. Finally, a decision tree classifier has been designed and 1500 words images of handwritten and printed forms (equal in number) are fed to the classifier to evaluate the performance of the present technique. An overall success rate of 96.80% is achieved.

Keywords: Handwritten and printed word classification; decision tree classifier; feature ranking; gray-level feature vector;

1. Introduction

Paperless offices and societies are new trends in this era of advanced technology. Document image analysis is one of the important steps towards achieving this. Almost every activity of the offices and societies in developing countries involves papers, which are in the form of petition files, application forms, reports, letters, account's details

* Corresponding author. Tel.:+91-9933053452.
*E-mail address:*malakarsamir@gmail.com

doi:10.1016/j.protcy.2013.12.428

etc. In most of the situations, we come across with numerous documents containing a mixture of handwritten and printed text. Railway reservation forms, bank cheques, memorandums, receipts etc are some instances of such documents. Interlacing of handwritten and printed text may be found at word level, text line level and paragraph level. For efficient handling of varieties of documents, digitization of the documents is a pressing need. But, the presence of both handwritten and printed texts makes this task a real challenge to the researchers.

An effective Optical Character Recognition (OCR) system would be helpful to solve these issues. But the next question may arise that whether a single OCR system will be sufficient for encoding both handwritten and printed text or not. As printed characters generally, have uniform shape and structure, encoding them is less challenging in comparison with their handwritten counterpart. This is because of the fact that shapes and structures of the handwritten characters vary from writer to writer. Even in a document written by a single writer these variations are sometimes are very distinct. So, the feature selection and classifier design are quite different for printed and handwritten character recognition. Even in the character segmentation and recognition steps, the characters of the said categories show different dimension of complexities.

Use of two different OCR systems leads to reduce the search space of the OCR and also facilitates the retrieval of handwritten and printed text documents. Therefore to implement a cross OCR system, separation of handwritten and printed text from said documents is very essential.

A few works [1-12] are found in the literature for automatic discrimination between handwritten and printed text, words and/ or characters from document images. The works as introduced may be classified in three different levels, namely, paragraph-level separation, text line-level separation and word-level separation. In the present work, word-level handwritten and printed text separation is carried out. The work, reported in [1], classified printed character, handwritten character, photograph, and painted image regions from mixed document image. The work uses two features namely, distributions of gradient vector directions and luminance levels using neural network as classifier. The work, described in [2], distinguishes machine-written and hand-written characters in a digitized image. This work uses 3-feature values: i) straightness of vertically oriented lines ii) straightness of horizontally oriented lines and iii) symmetry relative to different points. A feed-forward neural network has been used as classifier there.

In the work [3], X-Y cut algorithm is utilized to obtain the word block from a document image and then handwritten and printed words are classified using spatial feature and character block layout variance. In [4], machine-printed and hand-printed text classification schema based on statistical features for Bangla and Devnagari script has been described. The work [5] has proposed an algorithm based on regularity characteristics on the projection profile and the theory of hidden Markov models (HMMs) to distinguish between machine-printed and handwritten materials. Run-length histogram features and stroke density histogram features are applied in the work [6] to identify the handwritten/printed Chinese character as well as printed Chinese/English character.

Statistical texture features such as mean, standard deviation, smoothness, moment, uniformity, entropy and local range including local entropy has been introduced in [7] for word-level handwritten and printed text classification. As a classifier, they have used K-Nearest Neighbor (K-NN) learning mechanism. The work has been applied on scripts like Kannada, Telugu, Malayalam and Hindi. Another work, presented in [8], has devised a method for discriminating handwritten and printed text from document images based on shape features like area, perimeter, form factor, major and minor axes, roundness and compactness.

A two step approach is introduced in the work [9]. The steps are namely i) patch level separation and ii) pixel level separation. In the patch level separation, the entire document is classified into three different classes (machine printed text, handwritten text and overlapped text) using G-means based classification followed by a Markov Random Field (MRF) based relabeling procedure. In pixel level separation, a MRF based classification approach is performed to separate overlapped text into printed text and handwritten text using pixel level features. In the work [10] using run-length smoothing algorithm (RLSA) the extraction and classification of pseudo-lines and pseudo-words from document images are performed. Then the pseudo-words are used for classification purpose. The work used 4 different sets of features. They are i) morphological (local properties of pseudo-words such as height, width and pixel number), ii) connected component descriptors (11 descriptors as proposed in [11]), iii) pixel repartition (global descriptors like invariant Hu moments, variance of the projection profiles [12] and iv) other local properties such as run length, crossing count and bi-level co-occurrences, as described in [11]).

The problem of classification of handwritten and printed text in document image is not widely discussed compared to the other fields of OCR system. May be it seems simple to the researchers, though it is not, or may be its limited applications. In all of the works, described above, the word images are first binarized by any of the well-known binarization technique and then features are calculated on the binarized image. Binarization is an overhead for the technique, as some information may be lost during binarization process. Again, all the works have used some learning techniques for the classification process which also increase the computation time of the classification process. To the best knowledge of the authors, no work has considered gray-level intensity values of the handwritten/printed word images as feature value for the classification. So, in the present work, a simple and effective technique has been introduced which uses gray-level feature values of the images and a simple decision tree based classifier for separating the handwritten words from the printed ones.

2. Present Work

The work reported here is a classification problem which classifies the handwritten and printed word images in a document image where both types of the text are present. Firstly, the word images of these two aforesaid types (handwritten and printed both) are collected from different sources. Next, a 6-element feature set for each word image is designed and the ranking of the features is carried out. Depending upon the ranked feature, a tree like classifier has been designed and accordingly the unknown words are fed to evaluate the performance of the designed classifier. The block diagram of the proposed is shown in Fig. 1. All these steps are described here in brief in the following subsections.

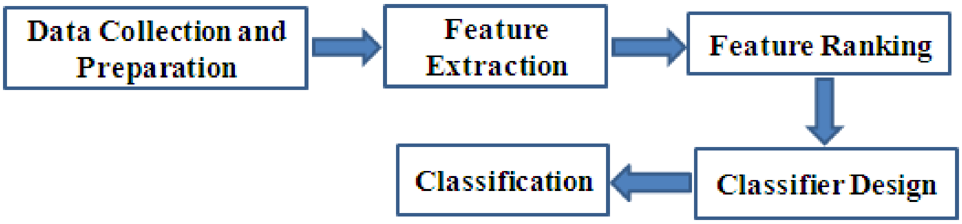


Fig.1. Block diagram of the proposed work

2.1. *Data collection and Preparation*

Some text documents mixed with both handwritten and printed forms are collected from different sources. These documents are scanned using flat-bedded scanner with 300 dpi resolution. Then 2000 words are cropped manually from these document images. Out of 2000 words 1000 are handwritten word and rest 1000 words are printed ones. 250 words of each category are chosen to construct the decision tree and rest 1500 (750 of each category) are kept for validating the decision tree classifier.

2.2. *Feature Selection*

It is already mentioned that printed word image and handwriting word image have different visual appearances i.e., different intensity distribution, so the selection of feature values in the present technique are based on gray-scale intensity values. A 6-tuple feature vector $F = \{Fi: i = 1, 2, \dots, 6\}$ has been designed for classifying handwritten and printed word images. A brief description of all the features, listed in Table 1, is given in the following subsections.

2.2.1. Mean pixel intensity

Let, $I(x, y) \in [0, 255]$ where $0 \le x < w$, $0 \le y < h$ is a word image with w and h are height and width respectively. The mean pixel intensity value (μ) a the word image is defined by,

$$\mu = \frac{\sum_{x=0}^{w-1}\sum_{y=0}^{h-1} I(x,y)}{w \times h} \tag{1}$$

The ink color, pressure of writing in the handwritten word images varies from writer to writer. Even the foreground pixel intensities do not remain closer to zero in most of the cases. That is, in general, the mean intensity values for a handwritten word images are more than printed word images and significantly different. A graphical representation of mean intensities of handwritten and printed word images have been shown in Fig. 2(a). From the figure it is quite obvious that the mean intensity of word image helps to classify the handwritten and printed word images.

Table 1.Features used for printed and handwritten word classification

Feature #	Feature Description
F1	Mean pixel intensity value of a word image
F2	Standard deviation of pixel intensities of a word image
F3	Otsu's threshold value of a word image
F4	Number of local maxima in pixel intensity histogram of word image
F5	Percentage of pixels belonging to upper quarter of the pixel intensities
F6	Percentage of pixels belonging to lower quarter of the pixel intensities

2.2.2. Standard deviation of pixel intensities of a word image

In statistics and probability theories, the standard deviation, represented by the symbol sigma(σ), shows how much variation or dispersion exists from mean data value. A low value of standard deviation indicates that the data points tend to be very close to the mean whereas the high value of standard deviation indicates that the data points are spread out over a large range of values. This concept is applied here i.e., the standard deviation of the pixel intensities is considered as another feature value.

The standard deviation (σ) of a word image is defined as follows

$$\sigma = \sqrt{\frac{1}{h * w}\sum_{x=0}^{h-1}\sum_{y=0}^{w-1}(I(x,y) - \mu)^2} \tag{2}$$

where μ is the mean of the intensity values of a word image $I(x, y)$.

A graphical representation of standard deviations of pixel intensities of handwritten and printed word images have been shown in Fig. 2(b).

2.2.3. Otsu's threshold value

As shown in Fig.2(a) that for some of the images the mean intensity values are overlapped. This so happen due to the changes in ratio of foreground and background pixels in a word image. From this observation, another feature value has been introduced here. This feature value is the threshold value for image binarization by Otsu's method [13]. Fig. 2(c) depicts a graphical representation for estimated threshold values by Otsu's method for different handwritten and printed word images. The graph shows that this feature also has significant contribution to the classification of handwritten and printed word images.

2.2.4. Number of local maxima in pixel intensity histogram of word image

The pixel intensity histogram is the count of number of pixels belonging to a particular intensity level of an image. For handwritten word document, intensity varies a lot and the number of local maxima in the pixel intensity histogram is larger than that of printed word images.

Let, Hist[i], $\mathrm{i} = 0, 1, \ldots, 255$ represents the histogram of the word image $\mathrm{I(x, y)}$, $0 \leq \mathrm{x} < w, 0 \leq y < h$ and $\forall x \in [0, \mathrm{h}) \wedge \forall \mathrm{y} \in [0, \mathrm{w}), \mathrm{I(x, y)} \in [0, 255]$. The existence of local maxima is considered using the following rules:

i. If $\mathrm{i} = 1, 2, \ldots, 254$ and $\mathrm{Hist[i-1]} < Hist[\mathrm{i}] < Hist[i+1]$
ii. If $\mathrm{i} = 0$ and $\mathrm{Hist[i+1]} < Hist[i]$ and
iii. If $\mathrm{i} = 255$ and $\mathrm{Hist[i-1]} < Hist[i]$

Depending upon these three rules, the total number of local maxima in a histogram of a word image is calculated. It is found that they strongly differ for handwritten and printed word images as shown in Fig. 2(d).

2.2.5. *Percentage of pixels belonging to upper quarter of the pixel intensities*

Generally, percentage of high intensity pixels (i.e., pixels with higher intensity values) for a handwritten word images is higher in amount than that of printed word images which is shown graphically in Fig. 2(e). Due to this fact, percentage of pixels of a word image belonging to upper quarter of intensity values is considered as a feature value here.

The quarter range (R) and percentage of pixels in an image $\mathrm{I(x, y)}$ belonging to upper quarter (H) is defined by

$$R = \frac{g_u - g_l}{4} \tag{3}$$

and $$H = \frac{N_h}{N} \times 100\% \tag{4}$$

where $\mathrm{g_u}$ = maximum gray level intensity value and $\mathrm{g_l}$ =minimum gray level intensity value and $\mathrm{N_h}$ = number of pixels with intensity value $\in [\mathrm{g_u} - \mathrm{R}, \ \mathrm{g_u}]$ and $\mathrm{N} = \ \mathrm{h} \times \mathrm{w}$.

2.2.6. *Percentage of pixels belonging to lower quarter of the pixel intensities*

Again percentage of pixels belonging to lower quarter of intensity values in a word image may vary significantly for handwritten version than that of the printed one as depicted in Fig. 2(f). Therefore, a similar mechanism has applied for estimating F5 i.e., percentage of pixels belonging to lower quarter (L) of intensity values in a word image and is defined by

$$L = \frac{N_l}{N} \times 100\% \tag{5}$$

where $\mathrm{N_l}$ = number of pixels with intensity value $\in [\mathrm{g_l}, \ \mathrm{g_l} + \mathrm{R}]$ and $\mathrm{N} = \ \mathrm{h} \times \mathrm{w}$.

2.3. *Feature ranking*

As already mentioned, depending upon the feature set F, a decision tree has been constructed to perform the said classification procedure. Precedence of better features is needed for better classification performance. So, the ranking of the features $Fi's, i = 1, 2, \ldots, 6$ is utmost required.

Let, Fij^k represents the i^{th} feature values for k^{th} word sample of j^{th} category. Here $j = 1$ represents the printed word and $j = 2$ represents the handwritten word.

The maximum (Fij_{max}) and minimum (Fij_{min}) feature values of i^{th} feature for j^{th} word category consisting of Njsamples can be defined as

$$Fij_{max} = \overset{Nj}{\underset{k=1}{max}} \{Fij^k\} \tag{5}$$

(6)

and $Fij_{min} = \min_{k=1}^{Nj} \{Fij^k\}$

The overlapping range of feature values of any feature Fi for both printed and handwritten words is called undecided range, $Fi_{undecided}$ $(i = 1, 2, \dots, 6)$.

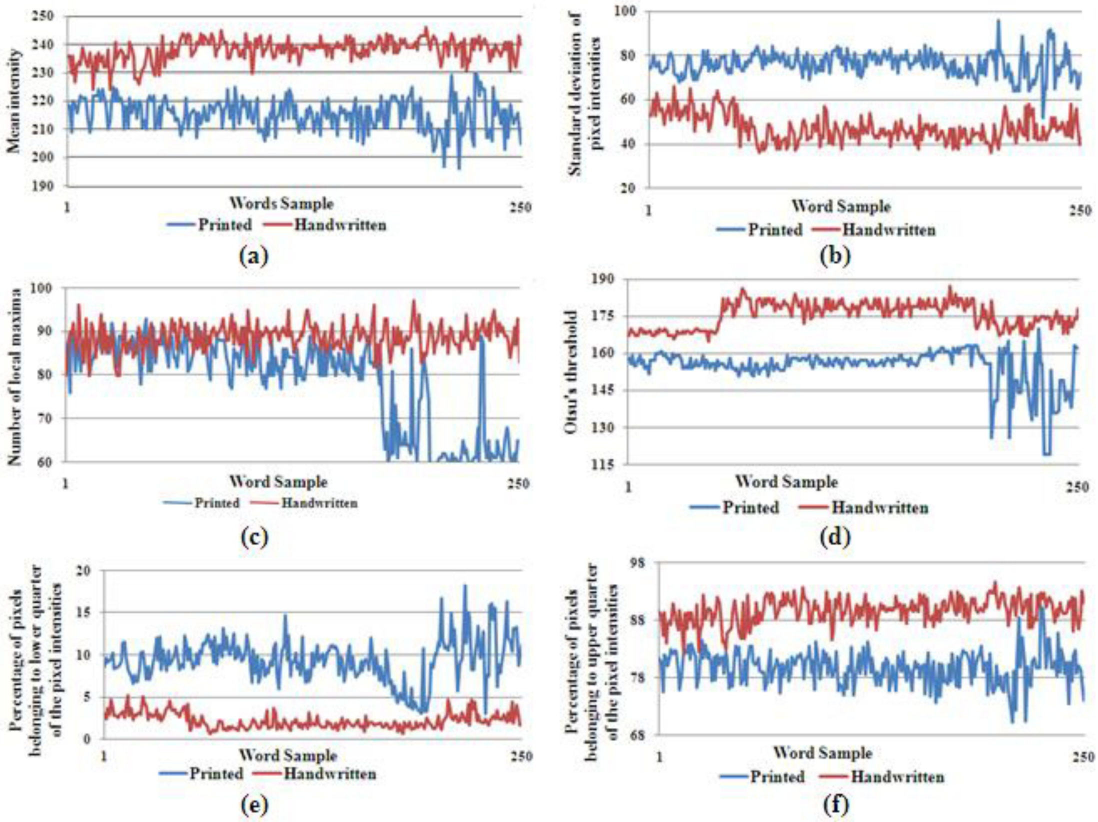


Fig. 2 (a-f). Comparative study on 6-gray level feature values for handwritten and printed word images represented graphically

$$Fi_{undecided} = \{X_i \mid X_i \in [Fi1_{min}, Fi1_{max}] \cap [Fi2_{min}, Fi2_{max}]\} = [a_i, b_i] \quad (7)$$

where $a_i = max\{Fi1_{min}, Fi2_{min}\}$ and $b_i = min\{Fi1_{max}, Fi2_{max}\}$

The spread (S_i) of $Fi_{undecided}$ and the ranking confidence (C_i) of a feature Fi can be defined by the following formulae

$$S_i = a_i - b_i \quad (8)$$

$$\text{and } C_i = \frac{N_W}{N} \times 100\% \quad (9)$$

where $N_W = |\{k: Fij^k \notin Fi_{undecided}\}|$ and $N = \sum_{j=1}^{2} Nj$

Depending upon this C_i values, Fi's are ranked.

2.4. *Decision tree classifier design*

After ranking the feature value a decision tree classifier has been designed for the classification of handwritten and printed word images. A generic decision tree is shown in Fig. 3. For better understanding the classifier, the required terminologies are.

Let $Ri = \{R1, R2, \dots, R6\}$ represent the ranked features. The range, termed as undecided range is formulated as

$$Ri_{undecided} = \{X_i | X_i \in [Ri1_{min}, Ri1_{max}] \cap [Ri2_{min}, Ri2_{max}]\} = [a_i, b_i] \tag{10}$$

The range of feature values of the ranked feature Ri represents only handwritten word images is termed as handwritten word zone i.e., $Hzone_{Ri}$ $(i = 1, 2, \dots, 6)$. This is expressed by the formula

$$Hzone_{Ri} = \{X_i | X_i \notin [a_i, b_i]\} = [c_i, d_i] \tag{11}$$

, where $c_i = \begin{cases} Fi2_{min}, & \text{if } Fi2_{min} \le a_i \\ 1 + b_i, & \text{if } Fi2_{min} \ge b_i \end{cases}$ and $d_i = \begin{cases} Fi2_{max}, & \text{if } Fi2_{max} \ge b_i \\ 1 - a_i, & \text{if } Fi2_{max} \le a_i \end{cases}$

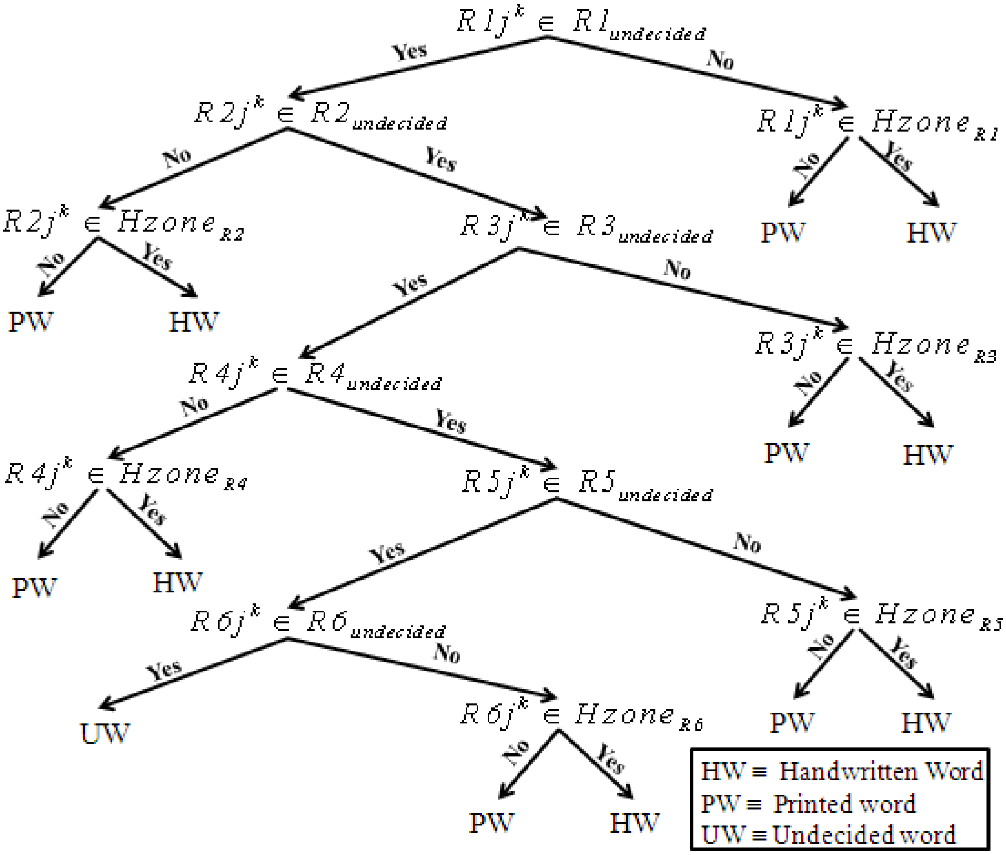


Fig. 3. Designed decision tree classifier

3. Results and Discussion

For experimental purpose, 2000 word images have been collected from different sources. Out of 2000 word images, 1000 words are handwritten words and rests are printed word images. Some of the sample word images are shown in Table 2. 500 word images of the two categories of images (equal in number) are used to construct the decision tree as described that 1500 word images are kept for validating the performance of the current technique. Table 3 describes the maximum and minimum feature values of each of the said 6 different features, obtained for handwritten and printed word images respectively experimenting on the training dataset.

Table 2. Some of the word images and their corresponding feature values

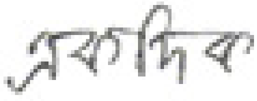
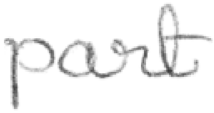
Handwritten words		Printed words	
Image	$\{Fi, i = 1,2, \dots ,6\}$	Image	$\{Fi, i = 1,2, \dots ,6\}$
	{238, 48, 148, 14, 89.36, 1.58}		{224, 70, 157, 84, 84.36, 6.93}
	{241, 36, 196, 76, 90.43, 0.26 }		{220, 75, 154, 80, 82.61, 10.14}

Table 3. Description of feature value, confidence and ranking of features

i	$Fi1_{min}$	$Fi1_{max}$	$Fi2_{min}$	$Fi2_{max}$	$Fi_{undecided}$	C_i (in %)	Modified Rank feature
1	196	230	224	246	[224, 230]	94.4	R_1
2	52	96	36	66	[36, 52]	86.2	R_3
3	119	170	165	187	[165,170]	83.2	R_4
4	59	93	80	97	[80, 93]	50.6	R_6
5	70	90	82	95	[82, 90]	63.4	R_5
6	3	18	1	5	[3, 5]	88.2	R_2

Depending upon the Ri ($\mathrm{i} = 1, 2, \dots , 6$) values $Ri_{undecided}$ and $Hzone_i$ have been confirmed and the corresponding values are fed into the decision tree as shown in Fig. 3. This transformed decision tree has been used as classifier in the present work. The overall dataset description and experimental success rate (in %) are tabulated in Table 4. Some of the correctly classified, misclassified and undecided word images are shown in Table 5.

4. Conclusion

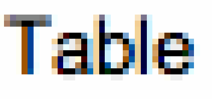
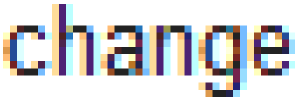
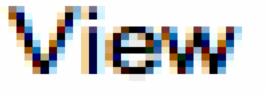
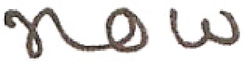
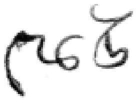
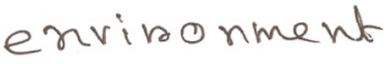
The paper reports a novel technique of classifying handwritten and printed word images collected from document images where both handwritten and printed texts are present. Here, a unique decision tree classifier is constructed for classification process with the help of 6 gray-level feature values specially designed for this purpose. The validity of the proposed scheme is evaluated 1500 word images. The present technique shows overall 96.80% success rate, which is satisfactory.

Though the feature values and a decision tree classifier produces reasonably good result, still the technique fails to categorize some of the handwritten word images. For that, introduction of more sophisticated feature values are required which is left as the future scope of the present work. Two-stage classification process i.e., extraction of feature values of gray images and binarized images may also be applied for achieving better classification result. Even some standard classifiers like Support Vector Machine (SVM), Multi Layer Perceptron (MLP) etc. may be considered for the enhancement of the technique.

Table 4. Overall classification statistics

Dataset	# of words	Classified as			Performance (%)		
		Handwritten	Printed	Undecided	Successful	Unsuccessful	Undecided
Handwritten	750	719	18	13	96.8	1.933	1.267
Printed	750	11	733	06			

Table 5. Sample classified, misclassified and rejected words by the present technique

Correctly classified word image	Misclassified word imag	Undecided word image
		
		

References

[1] S. Imade, S. Tatsuta and T. Wada, Segmentation and Classification for Mixed Text/Image Document Using Neural Network, in Proceedings of 2nd International Conference Document Analysis and Recognition, pp. 930-934 (1993)

[2] K. Kuhnke, L. Simoncini and Z. M. Kovacs-V, A system for machine-written and handwritten character distinction, in Proceedings International Conference Document Analysis and Recognition, pp. 811-814 (1995)

[3] K.C. Fan, L.S. Wang and Y.T. Tu, Classification of Machine-Printed and Handwritten Texts Using Character Block Layout Variance, in Pattern Recognition, vol. 31, No. 9, pp. 1275-1284(1998)

[4] U. Pal and B.B. Chaudhuri, Machine-Printed and Handwritten Text Lines Identification, Pattern Recognition Letters, vol. 22, nos. 3-4, pp. 431-441 (2001)

[5] J.K. Guo and M.Y. Ma, Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models, in Proceeding International Conference Document Analysis and Recognition, pp. 439-443 (2001).

[6] Y. Zheng, C. Liu and X. Ding, Single Character Type Identification, in Proceedings of SPIE Conference Document Recognition and Retrieval, pp. 49-56 (2002)

[7] M. Hangarge, K.C. Santosh, S. Doddamani, R. Pardeshi, Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents, in Proceedings of International Conference on Emerging Trends in Electrical, Communications and Information Technologies, Elsevier, pp. 215-221 (2012)

[8] P.L, Upasana, and M. Begum, Word Level Handwritten and Printed Text Separation Based on Shape Features, in International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol. 2, Issue 4 (2012)

[9] X. Peng, S. Setlur, V. Govindaraju, R. Sitaram, Handwritten text separation from annotated machine printed documents using Markov Random Fields, in International Journal on Document Analysis and Recognition pp. 1-16 (2013)

[10] A.Belaïd, K. C. Santosh, V. P. D'Andecy. Handwritten and Printed Text Separation in Real Document, in Machine Vision Applications, version 2 (2013)

[11] Y.Zheng, H.Li and D.Doermann, The segmentation and identification of handwriting in noisy document images, in Proceedings of DAS, pp. 95-105 (2002)

[12] R.Kandan, R. N.Kumar, K. R.Arvind, A. G.Ramakrishnan, A robust two level classification algorithm for text localization in documents, in the Proceedings of the Advances in visual computing, pp. 96-105 (2007).

[13] N. Otsu, A threshold selection method from gray-level histograms, in IEEE Transactions on Systems, Man and Cybernetics, vol. 9, no. 1, pp. 62-66 (1979).