Artif Intell Rev (2014) 42:253–273
DOI 10.1007/s10462-012-9337-z

Neural networks for document image preprocessing: state of the art

Amjad Rehman · Tanzila Saba

Published online: 29 April 2012

Abstract Neural network are most popular in the research community due to its generalization abilities. Additionally, it has been successfully implemented in biometrics, features selection, object tracking, document image preprocessing and classification. This paper specifically, clusters, summarize, interpret and evaluate neural networks in document Image preprocessing. The importance of the learning algorithms in neural networks training and testing for preprocessing is also highlighted. Finally, a critical analysis on the reviewed approaches and the future research guidelines in the field are suggested.

Keywords Learning algorithms · Preprocessing · Features selection · Slant correction · Line removal · Text differentiation

1 Introduction

Despite the wide use of electronic communication, paper document such as data entry forms, postal envelopes, and checks have central importance in our daily lives. As paper documents are cheap, reliable, secure for future reference, easily available and flexible in filling. Consequently, paper documents produced presently are more than ever before (Rehman and Saba 2011a). Additionally, most of the governments and private organizations use paper based documents to collect information. However, it is laborious, time consuming to collect handwritten information from forms and typed it into computers by human operator. Whereas, electronically processed documents are easy to process for searching, updating and for further

A. Rehman (✉)
College of Computer and Information Sciences, Al-Imam M. Saud Islamic University,
Riyadh, Kingdom of Saudi Arabia
e-mail: rkamjad@ccis.imamu.edu.sa

T. Saba
Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia,
81310 Skudai, Johor, Malaysia
e-mail: tanzilasaba@yahoo.com

processing. So digitizing the paper documents is worth. Accordingly, automation of this procedure has attracted intensive research work in this field (Rehman et al. 2008c, 2010, 2011). Unfortunately, even after decades of intensive research efforts in this domain, capabilities of the current OCR systems are still quite limited for printed text only and a small fraction of the data are entered into the computers automatically (Saba et al. 2011a). A complete solution to the automatic document layout analysis and its classification has not yet been well matured (Saba et al. 2011b). In this regards, preprocessing techniques perform a crucial role in the entire process of document layout analysis and classification (Rehman et al. 2009a). The organization of the rest of the paper is as such: Sect. 2 presents document image analysis techniques that are further divided into thresholding, noise removal, skew estimation and slant correction. Section 2.1 describes segmentation of documents into pages, lines, words and characters. Features extraction is detailed in Sect. 3; finally, conclusions are drawn in Sect. 4.

2 Pre-processing

Process of acquiring an electronic image of a paper based document using a flat-bed scanner, digital cameras or mobile phones is termed as document acquisition. In this process, some noise, skew and other unnecessary variations are unavoidable. Hence, preprocessing is mandatory in most of the document layout analysis and classification operations (Rehman et al. 2010, 2011). It basically enhances the actual image for suitable further analysis. The preprocessing may itself be broken into smaller tasks such as line removal, skew estimation and correction, base-line detection (upper and lower), smoothing and so on. Several methods have been proposed in the literature for estimating the above parameters. This section describes in detail afore mentioned pre-processing techniques.

2.1 Thresholding

Following the image acquisition stage that is normally attained via the use of a digital scanner one of the first pre-processing operations is thresholding. Accordingly, scanned image is stored in grayscale format, in which 0 (total absence, black) and 1 (total presence, white). Various shades of gray are represented between these two values. Many researchers have decided to convert the initial grey-level images into a less storage intensive format i.e. a binary (0 and 1), black and white format. It is argued whether recognition performed on features directly extracted from grayscale or from binary images produces the better result. The process of converting a grey-level image to a binary image is called thresholding or binarisation. This is the operation of selecting which elements of an image may be considered the background (white pixels) and which elements are to be considered the character itself (black pixels). Some threshold is usually used so that pixels of intensity over the threshold are marked as being background pixels and pixels of intensity less than threshold are considered to be part of the foreground image. Selecting an appropriate threshold has been the subject of active research for a number of years (Rehman et al. 2009b). However, Otsu (1979) is considered benchmark and is applied by several researchers. Hence, Otsu (1979) is implemented in this research to threshold IAM form images.

2.2 Line removal

In document images, printed lines are frequently used that overlapped with script. These lines are used to align the writer on the horizontal axis. Typical examples of such images

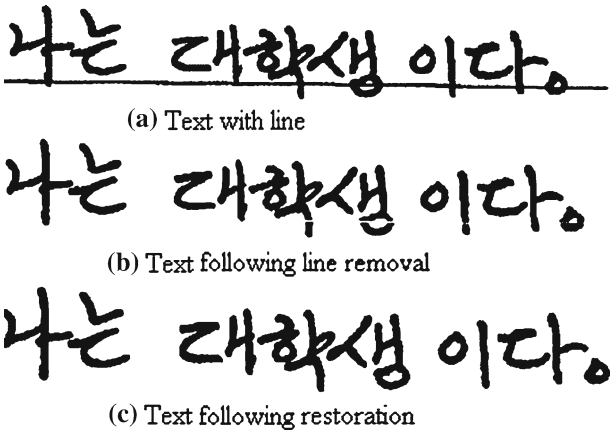

Fig. 1 Line removal from text (reprinted from Yong et al. 1997)

are bank cheque, receipts and payment slips. Additionally, hand drawn skewed lines make the dilemma more crucial. Among the others, text overlapping with underlines poses serious segmentation (if exists) and recognition problems, particularly when the documents must be filled manually by the writer according to the printed underlines. Furthermore, lines can be of different width and length; they may be broken and are connected to the handwritten text in many parts (Rehman et al. 2008a). Consequently, it is quite possible that some parts of the text overlap with the underline and therefore, may be deleted during line elimination (Rehman et al. 2009c). The detection and removal of these factors through preprocessing techniques can be helpful to reduce variability and to improve recognition rates (Saba and Rehman 2011).

Several approaches have been proposed for underline removal in the literature. Most of them detached underline from binarized image using dilation and erosion operators of the mathematical morphology (Rehman et al. 2009b). Dilation operator is applied until all the lines longer than a fixed threshold are removed from the underline region. On the other hand, this operator shatters the characters and therefore, it become difficult to recognize broken characters. As a result, erosion operator is applied to recover the lost parts of the characters as depicted in Fig. 1. However, broken characters could not restore correctly (Rehman et al. 2009c,d). Govindaraju and Srihari (1992) achieve underline removal by using the "good continuity criterion". Initially, in the image to detect smooth strokes, spine of the image is identified as the smooth stroke with maximum length, and is finally removed. However, this approach works out on thinned images and therefore, it requires a preliminary time consuming process. Yu and Jain (1996) propose a method for line removal and character restoration using Block Adjacency Graph representation of the input binary image. The horizontal form lines are located by finding long straight lines based on the block adjacency graph. Separation of form lines and character reconstruction are also implemented from this graph (Sulong et al. 2010).

Yoo et al. (1997) classify the various types of junction points at the point of contact or crossing over characters and line. After line removal, the junction points are detected and are restored based on their classification type. Koerich and Ling (1998) also detect and remove the lines using horizontal projection profile (HPP). The removed regions are rectified by checking the neighbours for every pixel that can be fitted into the erased line. Based on whether the neighbour pixels satisfy certain condition or not, the decision to leave the pixel

on or off is taken. In some algorithms such as proposed by Wang and Srihari (1991), broken characters are restored and send to the character recognizer. If result of the restored characters is wrong, restored characters are return to the restoration algorithm stage. In such methods, the processing time increases considerably because it has feedback paths in a recursive fashion. In addition, characters are sometimes recognized incorrectly such as 'h' and 'b' (Yong et al. 1997).

Blumenstein et al. (2002) introduce new pre-processing techniques for underline removal and restoration based on horizontal black pixel runs. It is assumed that word stroke thickness will be similar to the thickness of the underlines present in the word. However, this assumption is still not true in all cases particularly for printed documents/forms. Finally, authors report that underline removal and restoration do not perform well on some of the more erratic underlines that are present in some word images. Therefore, remainders of undetected underlines are removed manually to facilitate further processing.

Bai and Huo (2004) use strategies of connected component and bottom edge analysis to detect underline in printed text. Prior to removal of the detected underline, an OCR engine is employed to recognize and verify the input text line. However, the approach deals with removal of underline in printed text and failed in script line removal.

Recently, Arvind et al. (2007) detect multiple printed lines with varying thickness present in the word image using horizontal projection profile technique. Smashed characters are restored by using Bresenham line drawing algorithm (Foley et al. 1997). However, technique could not deal with restoration of script and skewed images. Recently, Rehman et al. (2011) presents a new approach to detect and remove unwanted printed line inherited in the text image at any position without character distortion to avoid restoration stage. The technique is based on connected component analysis and successfully dealt with underline, overlapped line and broken line removal from printed text and script promising results are reported. To conclude, common problems of reported techniques are.

(i) Computationally expensive as consists of two stages: line removal and restoration of smashed characters.
(ii) Deal with underline removal in printed text rather than line removal.
(iii) Cannot deal with line removal in script writing.
(iv) Cannot deal with skewed line removal in script writing.

Finally, Table 1 presents a critical review of line removal techniques available in the literature.

2.3 Slant estimation and correction

Slant correction is an indispensable technique for both holistic and analytical based script recognition (Rehman and Mohamad 2008). For slant correction, the crucial step is to find slant angle correctly. A slant is the clockwise angle between the vertical direction and the dominant direction of the vertical strokes. The aim of slant removal is to make the text invariant with respect to the slant angle. In the literature, several methods have been proposed to deal with this problem, mostly based on vertical projection profile. However, some approaches also utilize Wigner Ville distribution, cost function, Sobel operators, structure features using chain code to estimate the slant angle. Based on their fundamental strategy, slant estimation approaches are divided into three categories.

2.3.1 Projection profile-based approaches

Initially, Bozinovic and Srihari (1989), claim that vertical strokes of the word are main source of slant information. Therefore, for each selected stroke, centroid of the lower and upper half

Table 1 Critical review of line removal techniques in literature

Authors	Proposed methods	Comments
Serra (1982), Charles et al. (1988)	Mathematical morphology	Dilation operator is applied to remove underline but it damages the characters. Therefore, erosion operator is applied to restore the broken characters that increase computational complexity. Moreover, the technique deals with straight-line removal only.
Govindaraju and Srihari (1992)	Good continuity criterion	Approach is applicable to thinned images only.
Yong et al. (1997)	Morphological shape analysis	Deals with only printed straight lines and restoration of damaged characters carried out following line removal so processing time increased.
Dimauro et al. (1997)	Dynamic selection of structuring element	Does not seem to deal with skewed underline detection/ removal.
Blumenstein et al. (2002)	Foreground pixels analysis	Does not deal with real skewed line detection/ removal and thickness of line is determine by average stroke thickness that is not always true.
Rehman et al. (2010, 2011)	Connected component features	Does not need restoration of characters, as there is no damage of characters in line removal process.

retained and slope of the line connecting them is taken as a local slant estimate. Consequently, global slant estimate is calculated as an average of all local slant estimates and finally image is sheared at global slant estimate. Later, this idea leads several researchers and they use different ways to select the strokes (Rehman et al. 2010, 2011).

However, vertical projection profile based approaches shear the text for a discrete number of angles around the vertical orientation. For each of these images, the vertical projection profile is calculated. The profile giving the maximum variation is taken as the profile corresponding to the deslanted image. These methods use different criteria to select near-vertical strokes. The slopes of selected strokes are estimated from the contours. Marti and Bunke (2001) also analyze contours near vertical strokes, approximated by a set of straight lines l_i. The angle histogram thus obtains as follow.

$$h_{sl}(\alpha) = \sum_{\{i:\lambda_i=\alpha\}} |l_i|^2, \text{ slant estimate } \alpha_{sl} = \arg\max_{\alpha}(h_{sl}(\alpha)) \qquad (i)$$

where, h is the histogram and α is the angle.

Likewise, Shridhar and Kimura (1995) propose two more methods for slant estimation and correction. In the first one, the vertical projection profile is adopted while latter are based on chain code method.

Compared to the other approaches this method is efficient, but on the downside, it relies heavily on heuristics, hence is not robust. Additionally, such approaches require the detection of the edges of the characters and its accuracy depends on the characters included in the word (Rehman et al. 2008c).

The main disadvantage of methods based on Bozinovic and Srihari (1989) deslanting idea is that parameter tuning is needed which results in heavy experimental efforts to find the optimal point in the parameter space. Additionally, due to huge variety of handwriting styles, strokes, word dimensions, the estimated parameters might be sub-optimal in many cases. The process of slant correction also introduces noise in the contour of the image in the form of bumps and holes (Rehman 2010).

Few approaches evaluate a measured function of the image on a range of shear angles and select the angle with highest measured value. The measure is most often computed from the vertical histogram, based on the idea that deslanted writing has more intensive histogram i.e. with higher and more pronounced peaks than slanted writing. In this regard, Kavallieratou et al. (2003) propose a slant correction technique based on hybrid approach of Wigner–Ville distribution and projection profile technique. Wigner–Ville distribution is used to measure degree of variation through the different vertical projection profiles. The measure is repeated over shear transferred word images corresponding to different angles in an interval. The estimate with optimal angle is the slant angle. Finally, the hybrid slant correction approach is integrated in a complete image processing system (Kavallieratou et al. 2003). Despite, the approaches based on the optimization are relatively robust. However, it is computationally heavy since multiple shear transformed word images corresponding to different angles in an interval have to be calculated.

Vinciarelli and Luettin (2001) propose a similar deslanting technique based on hypothesis that the number of columns containing continuous stroke are maximum in a deslanted word. Therefore, a cost function is evaluated on multiple shear transformed word images. The angle with the maximal cost is taken as a slant angle estimate. On the downfall side, it has high computational cost because multiple shear-transformed word images corresponding to different angles in an interval have to be calculated (Dong et al. 2005). Several other proposed approaches are based on this idea such as El-Yacoubi et al. (1999), Cai and Liu (2000).

Some methods based on image convolution using Sobel edge operators are also available. Typically, these methods make a convolution of the image using vertical and horizontal Sobel kernels, where the gradient phase angle is calculated for every point of the image. A histogram of all angles is computed. In order to obtain the relevant angles (those that are close-to-vertical direction) a triangular of Gaussian filter, centered at 90°, applied to the histogram. The mean (or the most frequent) angle of the histogram is taken as the slant angle (Rehman and Saba 2011a).

To conclude, one of the major problems faced by all methods based on vertical projection profile is presented as below. For example, consider a binary image $I(i, j)$ with projection profile $P(j)$

$$\text{where, } P(j) = \sum_{i} I(i, j). \tag{ii}$$

If value 1 represents foreground black pixel then rows with higher horizontal projection profile values contain text. Hence, peaks of PP will be sharpened to provide localization of text lines. However, it is true for a deskewed and noise free image otherwise, text position will be unclear. Additionally, presence of frequent ascenders and descenders characters bring negative impact to PP based approaches.

Hence, projection profile based approaches are unable to distinguish among a scattered set of dots and a straight segment, if the number of black pixels is same. Therefore, a black pixel either from noise or from text line contributes equally in the projection. Due to these shortcomings projection profile based approaches do not seem to be reliable for the estimation of average slant angle (Saba et al. 2011c). Therefore, smoothing technique(s) are usually applied to remove contour noise. It has been mentioned previously, that some researchers employed skeleton format of the word image to normalize the stroke width. However, skeletonization operation for word recognition is still an issue having its own pros and cons.

2.3.2 Structure analysis

The second category of slant correction approaches explores structural features based on statistics of chain coded contours strokes to estimate slant angle. Marti and Bunke (2001) employ vertically oriented contour to estimate slant angle. Accordingly, they consider only horizontal black–white and white–black transitional pixels. The angle distribution of the writing's contour is accumulated in an angle histogram. Lines near to vertical parts are measured, to reduce the influence of horizontal contour. Likewise, Britto et al. (2000) and Kim (2003) utilize statistics of chain-coded stroke contours, weighted with their length, is considered as slant angle.

Chain code based approaches calculate an average slant angle for the whole word, which is sometimes, unsuitable for all characters in the word. Moreover, Britto et al. (2000) acknowledge that there are significant gaps between the average and the individual slant angles of component characters, which degraded the performance of analytical recognition systems.

Recently, Rehman et al. (2009a) propose slant estimation techniques based on structure features of first character. Accordingly, it is assumed that slant of whole word is almost equal to the slant of first character. An accuracy rate up to 96.74 % is reported. However, the technique fails if characters in the word have dissimilar slant.

2.3.3 Non-uniform slant correction

Lastly, approaches distinguish from all predecessors to correct non-uniform slant. The earlier techniques of slant removal shear a word (or bigger units) uniformly, i.e. by a single angle, hence cannot deal with non-uniform slanted characters in the words (Rehman et al. 2009a). On the other hand, it is an assumption that the slant angle fluctuates in a word due to various factors such as writer's habit, the inherent shape of each character, and writing position. This assumption raises the necessity to estimate local slant angles and to correct them non-uniformly. Linear searching techniques are employed to detect and correct non-uniform slant in the same word (Uchida et al. 2001). To apply different shear angles at different points within a word, one has to split the word up into intervals and shear each of those individually. To determine what intervals should be taken and by what angle to shear over each interval, a criterion is optimized which evaluates the sequences of intervals and angles simultaneously. However, optimal estimation is based on several constraints for local and the global validity of the local angles (Rehman and Saba 2011a).

Such methods have a lot of potential, since they can cope with variant-slanted words. The results are indeed promising, although there are more robustness issues, as the algorithm has greater freedom to make errors within a word. Furthermore, theoretical background and mathematical techniques are somewhat more processor demanding and therefore are computationally expensive.

2.4 Core-region (reference lines) detection

The estimation of core-region is also a critical task for the cursive handwriting recognition performance. It is a region between lower baseline and upper baseline. However, some strokes in a word image may extend above or below the core-region or main body of a handwriting sample, such letter components called ascenders and descenders respectively (Rehman et al. 2009c,d). Examples of letters that contain such strokes are 'f', 'j', 'g' etc. Hence, the core-region of a word image that does not contain ascenders and descenders bounded by an upper-baseline and baseline termed as core-region (Rehman and Saba 2011b).

Correct detection of core-region is of high importance as it serves for many purposes: such as the determination of height of the character for contextual information for example, to discriminate characters like "a" and "f" (Kurniawan et al. 2009a,b). Some researchers also use reference lines for skew correction such as Morita et al. (1999) perform skew correction via baseline detection. Mohamad et al. (2008) use core-region for segmentation of difficult cursive handwriting. Likewise, Cheng and Blumenstein (2005) perform ligatures analysis in the core-region and Verma (2002) use baseline to detect character contour for script segmentation.

In the literature, several methods for baselines detection are presented mainly based on projection profiles (PP). Cote et al. (1998) method is based on the entropy of the distribution (supposed to be lower when the word is desloped), while Vinciarelli and Luettin (2001) applied the Otsu method in order to find a threshold distinguishing between core region lines (above the threshold) and other lines. Finally, Blumenstein et al. (2002), Cheng and Blumenstein (2005) detect core-rigion by using horizontal density histogram (number of foreground pixels per line) proposed by Bozinovic and Srihari (1989). The baselines are erroneously detected because the "Peak Line" set incorrectly. This occurred because of a number of characters that contain large horizontal strokes such as the letter "t", "g", "h" etc. (Rehman et al. 2009a). Likewise, Morita et al. (1999) acknowledge that the horizontal density histogram is analyzed looking for features such as maxima and first derivative peaks, but these features are very sensitive to local characteristics and many heuristic rules are needed to find the actual core region lines.

Lee and Verma (2008a,b) propose an enhanced approach for core-region detection. Accordingly, upper baseline is located by measuring the distance from the upper-most pixel to the first foreground pixel for every single column. Number of transition is also measured vertically to exclude horizontal lines of few characters such as 'T', 'F' etc. Following calculation of transition features and distances, a search algorithm is applied to locate upper and lower baselines. Recently Rehman et al. (2009a) introduce quantile concept to detect core-region. An accuracy rate up to 98.62 % cases for lower baseline and 96.71 % for upper baseline detection is claimed.

2.5 Text differentiation in data entry forms

Despite of the extensive use of the electronic applications, paper documents still have central importance in our daily life. In this respect data entry forms are special documents typically used to collect or distribute data in both private and public sectors. Forms, particularly multilingual data entry forms are multifaceted and harder to deal than other documents because they contain not only printed text but also checkmarks, handwritings, and logos etc. Additionally, they vary in layout structure, style and size of text that makes text differentiation more critical. A typical filled-in form consists of two parts. One pre-printed machine text for the guidance of the user. Second are handwritten filled entries by the user. These two

parts often appear intermixed. On the other hand, the recognition methodologies of machine printed text are quite different from handwritten text. Therefore, to achieve optimum OCR rate in automatic form processing, first we have to distinguish machine printed text from handwritten text with high accuracy. Secondly, number of forms to be processed is usually very large, and there is a large variety of form types about several hundred to handle daily. So the processing speed is very important issue. The important technologies of auto form processing include type identification, data location and recognition. It is also still a research issue in the areas of document analysis (DA) and optical character recognition (OCR) (Saba et al. 2011d).

In the last few decades, a number of techniques are proposed for individual recognition of machine printed text and handwritten text independently. Accordingly, detailed reviews can be found in (Saba et al. 2011c).

It is evident from these surveys that nature of the machine printed text is quite different from the handwritten text. Accordingly, strategies for pre-processing, character segmentation, and features extraction for recognition are totally different from each other in all respective aspects. However, the classifying methods for machine-printed and handwritten characters have not been widely discussed in multilingual forms yet (Saba et al. 2010a; Bekhti et al. 2011). In the literature, many algorithms are proposed for discrimination of machine printed and handwritten text. However, most algorithms reported in the literature investigate typographic features for texture analysis. These typographic features are broadly divided into global and local features also termed as structural and statistical features (Saba et al. 2011c).

2.5.1 Text differentiation based on typographic features

Wang and Srihari (1991) search for intersections of line segments to distinguish printed and handwritten text. Imade et al. (1993) extract features from gradient vectors and luminance histogram to train neural network in order to differentiate printed and handwritten entries. The trained neural network is applied to separate a grey-level document image into photograph, painted image, handwritten Kanji and Kana character, and printed Kanji and Kana character regions. Srihari et al. (1994) report 95 % accuracy rate by extracting six symbolic features for Fisher's linear discriminant training classifier detailed as below.

(i) Standard deviation of connected component widths.
(ii) Standard deviation of connected component height.
(iii) Average component density.
(iv) Aspect ratio.
(v) Distinct different heights; and
(vi) Distinct different widths.

In the same way, Kuhnke et al. (1995) extract the direction features and symmetrical features of a single Roman character to classify into machine printed or handwritten using neural network. Liu et al. (1996) propose a method for discrimination of printed and handwritten entries by detecting form lines based on vertical density histogram. Belaid et al. (1995) apply Hough transform to detect and remove form lines. These lines are organized into a graph, and then search the graph for rectangles. The data inside the rectangles is supposed to be handwritten. However, many data entry forms do not use blank rectangles to be filled by the users.

Yuan et al. (1995) propose a new technique to detect handwritten fields and straight lines in data entry forms based on segmentation algorithm and adjacency graphs to detect possible entry fields in form images without entering text. Yu and Jain (1996) employ block adjacency

graph to extract the form frame. The detected form frame is dropped and the rest of the entries are supposed to be user-filled in. However, in this case layout of the forms needs to be known in advance.

Spitz (1997) initially divides scripts into two categories: first Asian scripts (Chinese, Japanese and Korean) and Roman script. This classification is based on the observation that upward concavities are distributed evenly along the vertical axis of Asian characters, but they also tend to appear at certain locations in Roman characters. Further distinctions among Asian scripts are made on the basis of character density. Hochberg et al. (1997) investigate a new technique for classification of thirteen languages scripts. They create a scale normalized cluster template for each script based on frequent characters or word shapes of this script, and then scripts are classified by comparing a subset of the document's textual symbols with these templates. However, the whole process is quite heavy and high storage demanded.

Fan et al. (1998) also employ typographical features for classification of machine printed and handwritten Roman and Chinese text only. They use spatial features and character block layout variance as the prime features in their approach to distinguish between machine printed and handwritten text. Accordingly, they divide character block into horizontal or vertical direction by analyzing the widths of the valleys of X and Y projection profiles of a text block image. Then, a reduced X–Y cut algorithm is utilized to obtain the base blocks from a text block image. An accuracy rate of 85 % is reported. However the approach is suitable only for Latin and Chinese text and heavily depends on the line, word, and character segmentation accuracy (Rehman and Saba 2011b). Xingyuan and Wen (1999) come out with a robust approach to detect rectangular fields and lines regardless of text or other markings with the assumption that rectangular fields consist of handwritten text. Choudhuri et al. (2000) propose a new approach for identification of Indian languages by combining Gabor filter based technique and direction distance histogram classifier. Their technique has considered Malayalam, Telugu, Bengali, Urdu, Hindi and English scripts. However, the approach demands high processing and storage (Chanda and Pal 2005).

Guo and Ma (2001) approach is based on the vertical projection profile of the segmented words. Accordingly, using hidden Markov model (HMM) as the classifier, classification accuracy of 97.2 % is reported. Guo and Ma (2001) combine statistical variations in projection profiles with HMM's to classify handwritten text from machine printed text. They hypothesize that machine printed text has a large number of regularities on the projection profile. However, this factor is missing in handwritten annotations because of variations in author, style, and environment. Chen and Lee (2001) present a gravitation-based algorithm for grouping and differentiating filled-in handwritten entries of form documents. Zheng et al. (2002) use aspect ratio and run-length histogram features to classify handwritten and printed Chinese characters and achieved promising results. However, they deal with characters discrimination and therefore, need character segmentation first that is the main bottleneck in segmentation phase particularly for cursive handwriting. Moreover, Pal et al. (2003) differentiate two Indian script (Bangla and Devnagari) based on statistical and structural features. They report an accuracy rate up to 98.6 %.

Nitz et al. (2003) detect text for mail facing and orientation purposes; however, no accuracy rate is mentioned for this specific task. Ma and Doermann (2003) propose a supervised multi-class classifier based on Gabor filters to classify scripts, font-faces, and font-styles (bold, italic, normal etc.). However, it is applicable for such applications where the classes are known in advance. Classification is performed at the word level (glyphs separated by white space) given training samples of each class. This method is applied to a variety of bilingual dictionaries to identify different scripts, and simultaneously, to classify Roman scripts into bold, italic and normal font-styles. An average accuracy up to 83.23 % is reported.

Kavallieratou and Stamatatos (2004) utilize structural features that usually help humans to discriminate printed from handwritten texts. In their opinion, the height of the printed characters is more or less stable within a printed text-line while the distribution of the height of handwritten characters is quite diverse. Thus, the ratio of ascender's and descender's height to main body's height would be stable in printed text and variable in handwritten. In a similar technique, Kavallieratou and Stamatatos (2004) propose an approach to discriminate machine printed and handwritten text based on bounding box. Accordingly all character blocks are bounded in box and structural features are extracted such as aspect ratio, area and projection peak. An accuracy rate for text separation up to 98.2 % is reported. However, these techniques need word and character segmentation that makes the problem more crucial (Rehman et al. 2010, 2011).

Raju et al. (2004) employ Gabor function based on filter bank to classify the text elements against all other kinds of clutter. They claim the technique to be working efficiently on camera captured images as well. Zheng et al. (2004) extract aspect ratio and run-length histogram features to feed a trained Fisher classifier in order to discriminate machine printed and handwritten text in noisy documents. Finally, a Markov random field-based (MRF) approach is applied to model the geometrical structure of the printed text, handwriting, and noise to rectify misclassifications. Although, results are good, yet they are high processor and memory demanded.

Chanda and Pal (2005) investigate an automatic approach for word wise identification of Devnagari, English and Urdu scripts in a single document. Zhou et al. (2006) perform identification of Bangla and English scripts in a single document based on the analysis of connected component profiles. However, their approach is language dependent. Xiuling et al. (2006) apply two-level regulated hit or miss transform (TLRHMT) feature of form to analyze and extract fields from a form image. They report accuracy rate up to 99.2 %. Arvind et al. (2006) conduct a comparative study of text/non-text separation based on various feature-classifier combinations. Features horizontal profile projection (HPP) and its two transformed versions, namely, Eigen and Fisher profiles are investigated in conjunction with nearest neighbour classifier, linear discriminate function, support vector machines and artificial neural networks. Some of these combinations perform text separation with accuracy more than 90 %.

Similarly, Nikolaidis and Strouthopoulos (2008) identify marks based on contour detection. The principal axes of marks are determined using PCA (principal component analyzer) and a nearest neighbour technique is used to find the shortest distances between marks. A feature vector is formed based on mark dimensions and distances between them, which is then fed into SOFM (self organizing feature map) in order to divide the marks into homogeneous clusters. A set of fuzzy rules is formed using all cluster weights and variances. Finally, using a fuzzy classification scheme each mark is divided into character or non-character pattern. However, no accuracy rate is reported.

Koyama et al. (2008a) propose a two steps spectrum-domain local fluctuation detection (SDLFD) method to discriminate machine printed and hand printed characters. First, they transform local region of a document image into frequency domain to extract feature values including fluctuations caused by handwriting. Next, the extracted features values are fed to a trained multilayer perceptron (MLP) to get likelihood of handwriting. An accuracy rate up to 97 % is reported. However, they deal with character distinction only.

Vijaya and Padma (2009) perform language identification for Kannada, Hindi and English text lines from printed documents. The approach is based on the analysis of the top and bottom profiles of individual text lines. Although high classification accuracy is reported, however, technique deals with only machine printed text.

Recently, Saba (2012a), introduced a combination of structure and statistical features to differentiate text into script and printed text. The statistical features include number of strokes below baseline and structure features consist of angle features in four quadrants. An accuracy rate up to 91 % for horizontal text block and 94.71 % for horizontal text line.

2.5.2 Text differentiation based on human vision mechanism

The methods mentioned in previous section need character or text lines from documents in pre-processing stage to extract typographical features for text discrimination. Although promising results are reported using texture analysis yet distinction rate is affected by accuracy of the line and character segmentation (Koyama et al. 2008b). Koyama et al. (2008b) suggest feature value E to differentiate handwritten character from machine-printed character. Mechanism of human vision inspires the feature value. The anticipated method makes use of power spectrum attains by locally computed two-dimensional Fourier transform of a document image. Definition of the feature value E is fluctuation-to-total power ratio.

Finally, Table 2 exhibits a comparison of the text differentiation approaches in the state of art.

To conclude, all existing text discrimination techniques including the methods mentioned above are either local or global approaches. The local approaches analyze a list of connected components (like line, word and character) in the document images to identify the text. However, these components are available only after line, word and character. In contrast, global approaches employ analysis of regions comprising at least two lines and hence do not require fine segmentation. Consequently, the language classification task is simplified and faster with the global rather than the local approach. However, in practice it is not possible to apply the global approach for the types of documents where one paragraph or one line itself is composed of more than one format (handwritten, machine printed). For such types of documents where the text type differs at paragraph and/or line level, it is necessary to apply local approaches.

Table 2 Text differentiation: a comparison

Author(s)	Methodology	Remarks
Vijaya and Padma (2009)	Based on analysis of the top and bottom profiles of individual text lines	Deals with only printed text
Koyama et al. (2008a)	Two steps spectrum-domain local fluctuation detection (SDLFD) method	Deal with character distinction only and needs training and training data. Computationally expensive.
Nikolaidis and Strouthopoulos (2008)	Based on SOFM (self organizing feature map)	Needs training/training data Time consuming.
Arvind et al. (2006)	Nearest neighbour classifier	Needs training/training data Time consuming Accuracy rate 90 %
Saba (2012a)	Hybrid features (statistical and structure features)	No training, training data, fast. Accuracy rate above 91 %

Finally, texture analysis is performed by most of the proposed techniques that are available in the literature. Additionally, different types of texture analysis tools, especially two-dimensional Fourier Transform, two-dimensional wavelet transform and multi-channel Gabor filter (Ma and Doermann 2003) are also used. Nevertheless, the entire process becomes computationally expensive with the use of these tools. Additionally, these tools need training data that varies from one language to another.

3 Feature extraction

The purpose of feature extraction is to achieve most relevant and discriminative features to identify a symbol uniquely (Saba et al. 2010b; Elarbi-Boudihir et al. 2011; Rahim et al. 2012a; Sulong et al. 2009). In OCR applications, extracted features are used to distinguish between all existing character classes. The research community is agreed that discriminative feature plays an important role in successful recognition of printed and cursive characters (Kurniawan et al. 2011; Rahim et al. 2011). Accordingly, many feature extraction technique are proposed and investigated in the literature that may be used for numeral and character recognition. Consequently, recent techniques show very promising results for separated handwritten numerals recognition (Rahim et al. 2012b; Haron et al. 2012), however the same accuracy has not been attained for cursive character classification (Saba and Rehman 2012). It is mainly due to ambiguity of the character without context of the entire word (Rehman and Saba 2012a,b; Phetchanchai et al. 2010). Second problem is the illegibility of some characters due to nature of cursive handwriting, distorted and broken characters (Rehman et al. 2008b; Haron et al. 2010). Finally, the segmentation process may cause some irregularities depending on the adopted approach (Kurniawan et al. 2009a,b, 2010). Research community has been trying to surmount above-mentioned problems in two ways. Firstly, exploration of features best for recognition. Secondly, search of different classification schemes and their fusion (Haron et al. 2011; Harouni et al. 2012).

According to Suen (1986), there are two main categories of features: statistical features and structure features. Statistical features derive from statistical distribution of every point in a character matrix such as moments, histograms, profile projection and zoning (Rehman and Saba 2012a). Statistical features are also known as global features as they are usually extracted and averaged in sub-images such as meshes. Initially, statistical features are developed to recognize machine printed characters (Vamvakas et al. 2007). While, structural features are based on geometric and topological features of characters such as contours, loops, end points (Saba et al. 2009). In this regard, Trier and Jain (1995) present a detailed review of feature extraction methods for offline isolated character recognition such as template matching, deformable templates, zoning, contour profile, profile projection, geometric moments invariants, Zernike moments, Fourier descriptors, Spline curve estimation. The methods are applicable to gray level character images, binary character images, thinned character images, character contours and character graphs. The statistical features in its simplest form use all points in the character matrix of binary image that contains black and white pixels only. Initially, all extracted features are employed to template matching, where similarities are judged from measured distance between unknown pattern and stored patterns. The minimum distance between unknown pattern and stored patters is the best match (Rehman and Saba 2012b). On the other hand, uses of all that features create dimensionality problems commonly known as curse of dimensionality (Saba et al. 2011b). Therefore, to overcome this problem, researchers explore statistical distribution of points in a character matrix. In statistical domain, five methods are reported in the literature.

3.1 Moments

Initially Hu (1962) introduces moments features for pattern recognition, commonly known as moments invariants as features are invariant to rotation, scale and translation. Later, the research community has extensively used invariants with less or more modification such as raw moments and central moments. Raw moments are used as a coordinate function for each point in the image. Suen (1986) and Impedovo et al. (1991) calculate central moments by taking distance of points from center of gravity. It is demonstrated that central moments enhance recognition accuracy and are invariant to the translation of the images in comparison of raw moments.

3.2 Zoning

In this method of feature extraction, character matrix is divided into several small zones. These zones are further manipulated from different aspects to extract features such as density of zones, transitions of foreground to background and vice versa. Projection: where the 2D image is represented as one-dimensional and features are computed accordingly (Kim et al. 2000). Gader et al. (1997) propose feature extraction technique based on transition information for character classification. Their technique calculates and locates transition of pixels from background to foreground and vice versa horizontally and vertically. Recently, Vamvakas et al. (2007), propose a hybrid feature extraction scheme for isolated Greek character recognition. The first scheme calculates density of each zone composing character image and second computes projection area of left, right, up and down character's profile. The extracted features are fused to create a single feature vector for character classification. Finally, support vector machine is employed to classify characters taken from IAM benchmark database; an accuracy rate of 92.91 % is reported.

3.3 Crossing and distances

Few researchers have also employed crossing and distances features. Number of times an image crossed vector at certain angles such as $0°, 45°, 90°$ etc and distance of character boundaries from a bounding box. Kim et al. (2000) extract two sets of features using crossing and distances. The first set consists of transition features that are calculated from rows and columns of character image and the second set is composed of distances of the first image pixel detected from the upper and lower boundaries along the vertical lines and from left and right boundaries along horizontal lines (Saba et al. 2011b).

(i) Characteristic loci: In this type of features, horizontal and vertical vectors are crossed from each background pixel and count crossing of line segment in the horizontal and vertical direction.
(ii) n-Tuples: This feature extraction scheme simply uses occurrence of all black and white pixels in the image.

On the other hand, in literature, structural features are also investigated to represent characters with high tolerance to distortions and style variations. In this domain, researchers have explored extensively geometrical and topological features of character image (Saba et al. 2010b). In this regard, Arica and Yarman-Vural (2001) also present a comprehensive review on structural features information. It includes all structural features such as: maxima and minima, shape, cross points, branch points, strokes and their direction, cusps below and above a threshold, angular change, degree of curvature, nature of the starting and ending points, their coordinates, the distance and the primitive features such as line segments, intersection of line

segment, length of line segment, convex polygons, and loops, etc. Moreover, Trier and Jain (1995) emphasis that contour can also be used as features such as height ratio,width of the character, vertical and horizontal profile ratio, location of maxima and minima in profiles. However, popular techniques for structural features extraction are coding and graphs. In the former technique, strokes of the character are mapped in 2D parameter space and in the later, the character is divided into a set of topological fragments such as strokes, loops, cross points etc then these fragments are represented by relational graphs (Arica and Yarman-Vural 2001). Again, there are two types of image representation by graphs. The first explores relationships between strokes and edges and second, uses coordinates of the character shape (Haron et al. 2012).

A number of techniques extract features from character's contours. Kimura and Shridhar (1991) divide contour profile into two halves and discrete function of each half is approximated to extract features. Yamada and Nakano (1996) explore direction histogram in character image to extract features. A multi-template based strategy with clustering feature is adopted to recognize segmented characters. Likewise, Kimura et al. (1997) evaluate features by calculating local histograms based on chain code information for segmented character classification. Krzyyzak et al. (1990) extract features from inner and outer contours of characters: simple topological features extracted from the inner contours and fifteen Fourier descriptors are extracted from the outer contours. Oh and Suen (1998) extract two feature set based on distance transformation and directional distance distribution (DDD). In the first feature set, distance of each white pixel to the nearest black pixel in the character image is calculated without character skeletonization. The second feature set composes of information encoding both black/white and directional distance distributions. Additionally, a new method of map tiling is introduced and is applied to the DDD feature to improve its discriminative ability. All experiments are carried out on three different sets of characters consisting of numerals, English letters, and Hangul letters. Promising results reported to confirm the best combination of DDD feature and the map tiling. Blumenstein et al. (2004, 2007) and Verma et al. (2004) use directional features extracted from character contours. The technique replaces foreground pixels of character contours with suitable direction values. Finally, image is divided into windows to extract features. Likewise, Mitrpanont and Limkonglap (2007) also analyze contours of Thai characters to capture movement of features for Thai character recognition.

Other studies by some researchers have also integrated these complementary features (statistical and structural) to investigate properties of character in order to improve recognition accuracy (Saba et al. 2010b). Camastra and Vinciarelli (2003) propose hybrid feature extraction technique for character recognition. Structural features derive from foreground pixel density and directional information while statistical features include character's fraction below baseline and character's width/ height ratio. Recently, Vamvakas et al. (2007) integrate two types of statistical features: density features and profile projection features to recognize offline Greek characters. They emphasized that hybridization brought better recognition accuracy as compare to individual scheme.

Fourier descriptors are also explored by research community for character recognition. Shridhar and Badreldin (1984) combine topological features with Fourier descriptors to enhance character recognition performance. Likewise, Trier and Jain (1995) use Fourier descriptors to represent skeleton of characters. On the other hand, Fourier descriptors have some disadvantages: above all these features are not helpful to detect spur on the boundary of characters and therefore, cannot distinguish between O and a Q (Kurniawan et al. 2011). However, this feature could be advantageous to filter noise on character's boundary. Table 3

Table 3 Features extraction methodologies: a comparison

Author(s)	Methodology	Accuracy (%)	Remarks
Saba (2012b)	Hybrid statistical and structure features	90.18	Suitable for touched script.
Rehman (2010)	Fused statistical features	91.38	Several features are extracted. Touched pattern are out of scope.
Vamvakas et al. (2007)	Hybrid feature extraction scheme	85.08	Only for isolated Greek character
Blumenstein et al. (2007)	Modified direction features	89.01	Isolated characters. Touched pattern out of scope.
Blumenstein et al. (2004)	Directional features	70.22 (lower) 84. 83 (upper)	Standard segmented characters are taken from benchmark CEDAR database and low accuracy
Camastra and Vinciarelli (2003)	Hybrid feature extraction technique	84.52	High processing demand technique
Singh and Hewitt (2000)	linear discriminant analysis-classifier	67.30	Low accuracy rate

presents latest features extraction methodologies for segmented character recognition in the state of art.

4 Conclusion

This paper has presented state of the art on document image preprocessing techniques using neural network. Literature reviewed in this paper emphasis that all existing preprocessing techniques have some inherent problems. Regarding text differentiation, performance of existing techniques depends heavily on lines, words and characters segmentation accuracy as well as orientation of characters or character blocks. Text that is not horizontally located on the page is usually discarded. Additionally, some techniques need training/training data and in most of the cases typical training data is unavailable. Time spend on training artificial intelligent tools is another issue (Rehman and Saba 2013). Finally, in practical, due to handwriting fluctuations performance of the existing text distinction methods is significantly low (Norouzi et al. 2012). Hence, efficient and training free text differentiation techniques are desired to differentiate text into machine printed and script categories accurately (Mohamad et al. 2012). Finally, to increase the character recognition accuracy, novel features extraction techniques are indispensible.

Acknowledgments This research work is partially supported by TWAS fellowship cycle (2010) and Higher Education Commission of Pakistan (partial support scheme). The authors are sincerely thankful to colleagues for guidance about the material and organization of this research work.

References

Arica N, Yarman-Vural FT (2001) An overview of character recognition focused on off-line handwriting. IEEE Trans Syst Man Cybernet Part C Appl Rev 31(2):216–233

Arvind KR, Kumar J, Ramakrishnan AG (2007) Line removal and restoration of handwritten strokes. In: International conference on computational intelligence and multimedia applications, pp 208–214
Arvind KR, Pati PB, Ramakrishnan AG (2006) Automatic text block separation in document images. In: Proceedings of 4th international conference on intelligent sensing and information processing, pp 53–58
Bai Z-L, Huo Q (2004) Underline detection and removal in a document image using multiple strategies. In: Proceedings of the 17th international conference on pattern recognition (ICPR04), vol 2, pp 578–581
Bekhti S, Rehman A, Al-Harbi M, Saba T (2011) AQuASys an Arabic question-answering system based on extensive question analysis and answer relevance scoring. Inf Comput Int J Acad Res 3(4):45–54
Belaid Y, Belaid A, Turolla E (1995) Item searching in forms: application to French tax forms. In: Proceedings of 3rd international conference on document analysis and recognition, pp 744–747
Blumenstein M, Cheng CK, Liu XY (2002) New preprocessing techniques for handwritten word recognition. In: Proceedings of second international conference on visualization, imaging and image processing, ACTA. Press, Calgary, pp 480–484
Blumenstein M, Liu XY, Verma B (2004) A modified direction feature for cursive character recognition. In: Proceedings of the international joint conference on neural networks, Budapest, Hungary, pp 2983–2989
Blumenstein M, Liu XY, Verma B (2007) An investigation of the modified direction feature for cursive character recognition. Pattern Recognit 40:376–388
Bozinovic RM, Srihari SN (1989) Off-line cursive script word recognition. IEEE Trans Pattern Anal Mach Intell 11(1):68–83
Britto AS, Sabourin JR, Latherier E, Bortolozzi F, Suen CY (2000) Improvement in handwritten numeral string recognition by slant normalization and contextual information. In: Proceeding of seventh international workshop on frontiers in handwriting recognition, pp 323–332
Cai J, Liu Z-Q (2000) Off-line unconstrained handwritten word recognition. Int J Pattern Recognit Artif Intell 14(3):259–280
Camastra F, Vinciarelli A (2003) Combining neural gas and learning vector quantization for cursive character recognition. Neuro-computing 51:147–159
Chanda S, Pal U (2005) English, Devanagari and Urdu text identification. In: Proceedings of international conference on document analysis and recognition, pp 538–545
Charles R, Giardina, Edward R, Dougherty (1988) Morphological methods in image and signal processing, Prentice Hall, Inc., p 217
Chen J-L, Lee H-J (2001) Field data extraction for form document processing using a gravitation-based algorithm. Pattern Recognit 34:1741–1750
Cheng CK, Blumenstein M (2005) The neural based segmentation of cursive words using enhanced heuristics. In: Proceedings of the eighth international conference on document analysis and recognition, vol 2, pp 650–654
Choudhuri S, Harit G, Madnani S, Shet RB (2000) Identification of scripts of Indian languages by combining trainable classifiers. In: Proceedings of international conference on vision, graphics and image processing
Cote M, Lecolinet E, Cheriet M, Suen CY (1998) Automatic reading of cursive scripts using a reading model and perceptual concepts—the PERCEPTO system. Int J Doc Anal Recognit 1(1):3–17
Dimauro G, Impedovo S, Pirlo G, Salzo A (1997) In: Removing underlines from handwritten text: an experimental investigation. Downton AC, Impedovo S (eds) Progress in Handwriting Recognition. World Scientific Publishing, pp 497–501
Dong J-X, Dominique P, Krzyyzak A, Suen C-Y (2005) Cursive word skew/slant corrections based on radon transform. In: Proceedings of the eighth international conference on document analysis and recognition, pp 478–483
Elarbi-Boudihir M, Rehman A, Saba T (2011) Video motion perception using operation Gabor filter. Int J Phys Sci 6(12):2799–2806
El-Yacoubi A, Gilloux M, Sabourin R, Suen CY (1999) An HMM-based approach for on-line unconstrained handwritten word modeling and recognition. IEEE Trans Pattern Anal Mach Intell 21(8):752–760
Fan KC, Wang LS, Tu YT (1998) Classification of machine-printed and handwritten texts using character block layout variance. Pattern Recognit 31(9):1275–1284
Foley JD, Dam AV, Feiner SK, Hughes JF (1997) Computer graphics: principles and practice in C, 2nd edn. Addison-Wesley, Pearson Education, Reading, MA
Gader PD, Mohamed M, Chiang JH (1997) Handwritten word recognition with character and inter-character neural networks. IEEE Trans Syst Man Cybern Part B Cybern 27:158–164
Govindaraju V, Srihari SH (1992) In: Separating handwritten text from interfering strokes. Impedovo S, Simon JC (eds) From pixels to features III—frontiers in handwriting recognition. North-Holland Publication, Amsterdam, pp 17–28
Guo JK, Ma MY (2001) Separating from machine printed text using hidden Markov model. In: Proceedings of the international conference on document analysis and recognition, pp 439–443

Haron H, Rahim S, Rehman A, Saba T (2010) Curve length estimation using vertix chain code. Int J Comput Sci Eng 2(6):2110–2113
Haron H, Rehman A, Adi DIS, Lim SP, Saba T (2012) Parameterization method on B-spline curve. Math Probl Eng 2012. doi:10.1155/2012/640472
Haron H, Rehman A, Wulandhari LA, Saba T (2011) Improved vertex chain code algorithm for curve length estimation. J Comput Sci 7(5): 736–743. doi:10.3844/jcssp.2011.736.743
Harouni M, Rahim MSM, Mohamad D, Rehman A, Saba T (2012) Online cursive Persian/Arabic character recognition by detecting critical points. Int J Acad Res 4(2):208–213
Hochberg J, Kelly P, Thomas T, Kerns L (1997) Automatic script identification from document images using cluster-based templates. IEEE Trans Pattern Anal Mach Intell 19(2):176–181
Hu MK (1962) Visual pattern recognition by moment invariants. IRE Trans Inf Theory 8:179–187
Imade S, Tatsuta S, Wada T (1993) Segmentation and classification for mixed text/image document using neural network. In: Proceedings of 2nd international conference on document analysis and recognition, pp 930–934
Impedovo S, Ottaviano L, Occhinegro S (1991) Optical character recognition—a survey. Int J Pattern Recognit Artif Intell 5:1–24
Kavallieratou E, Stamatatos S (2004) Discrimination of machine-printed from handwritten text using simple structural characteristics. In: Proceedings of the 17th international conference on pattern recognition (ICPR'04), pp 437–440
Kavallieratou E, Sgarbas K, Fakotakis N, Kokkinakis G (2003) Handwritten word recognition based on structural characteristics and lexical support. In: Proceedings of seventh international conference on document analysis and recognition, vol 1, pp 562–566
Kim D (2003) Slant correction of handwritten strings based on structural properties of Korean characters. Pattern Recognit Lett 12:2093–2101
Kim JH, Kim KK, Suen CY (2000) Hybrid schemes of homogeneous and heterogeneous classifiers for cursive word recognition. In: Proceedings of 7th international workshop on frontiers in handwriting recognition, pp 433–442
Kimura F, Shridhar M (1991) Handwritten numerical recognition based on multiple algorithms. Pattern Recognit 24:969–983
Kimura F, Kayahara N, Miyake Y, Shridhar M (1997) Machine and human recognition of segmented characters from handwritten words. In: Proceedings of 4th international conference on document analysis and recognition (ICDAR '97), pp 866–869
Koerich AL, Ling LL (1998) A system for automatic extraction of the user-entered data from bank checks. In: Proceedings of international symposium on computer graphics, image processing and vision, 270–278
Koyama J, Kato M, Hirose A (2008a) Distinction between handwritten and machine printed characters with no need to locate character or text line position. In: Proceedings of international joint conference on neural networks (IJCNN'08), pp 4044–4051
Koyama J, Kato M, Hirose A (2008b) Handwritten character distinction method inspired by human vision mechanism. In: LNCS Springer, pp 1031–1040
Krzyyzak A, Dai W, Suen CY (1990) Unconstrained handwritten character recognition using modified back propagation model. In: Proceedings of international workshop frontiers in handwritten recognition, April 1990, pp 145–153
Kuhnke K, Simoncini L, Kovacs ZM (1995) A system for machine-written and hand-written character distinction. In: International conference on document analysis and recognition, vol 2, pp 811-814
Kurniawan F, Rahim MSM, Daman D, Rehman A, Mohamad D, Mariyam S (2011) Region-based touched character segmentation in handwritten words. Int J Innov Comput Inf Control 7(6):3107–3120
Kurniawan F, Rehman A, Mohamad D (2009a) Contour vs non-contour based word segmentation from handwritten text lines. An experimental analysis. Int J Digit Content Technol Appl 3(2):127–131
Kurniawan F, Rehman A, Mohamad D (2009b) From contours to characters segmentation of cursive handwritten words with neural assistance. In: Proceedings of IEEE international conference on instrumentation, communications, information technology and biomedical engineering (ICICI-BME), pp 12–18
Kurniawan F, Rehman A, Mohamed D, Mariyam S (2010) Self organizing features map with improved segmentation to identify touching of adjacent characters in handwritten words. In: Proceedings of IEEE ninth international conference on hybrid intelligent systems, 2009. HIS '09, China, pp 475–480
Lee H, Verma B (2008a) A novel multiple experts and fusion based segmentation algorithm for cursive handwriting recognition. In: Proceedings of the international joint conference on neural networks (IJCNN'08), pp 2994–2999
Lee H, Verma B (2008b) Over-segmentation and validation strategy for offline cursive handwriting recognition. In: Proceedings of the international conference on intelligent servers, sensor networks and information processing, pp 91–96

Liu K, Suen CY, Nadal C (1996) Automatic extraction of items from cheques images for payment recognition. In: International conference on pattern recognition, pp 798–802
Ma H, Doermann D (2003) Gabor filter based multiclass classifier for scanned document images. In: Proceedings of 7th international conference on document analysis and recognition, pp 968–972
Marti U, Bunke H (2001) Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. Int J Pattern Recognit Artif Intell 15(1):65–90
Mitrpanont JL, Limkonglap U (2007) Using contour analysis to improve feature extraction in Thai handwritten character recognition systems. In: Seventh IEEE international conference on computer and information technology, CIT 2007, pp 668–673
Mohamad D, Rehman A, Kurniawan F (2008) A new approach for segmenting difficult cursive handwritten words from benchmark database. In: Proceedings of 4th international conference on information and communication technology and systems (ICTS) vol 1, pp 17–21
Mohamad I, Rahim MSM, Bade A, Rehman A, Saba T (2012) Enhancement of the refinement process for surface of 3D object. J Am Sci **8**(4):358–365
Morita M, Facon J, Bortolozzi F, Garnes S, Sabourin R (1999) Mathematical morphology and weighted least squares to correct handwriting baseline skew. In: Proceedings of the international conference on document analysis and recognition, vol 1, pp 430–433
Nikolaidis A, Strouthopoulos C (2008) Robust text extraction in mixed-type binary documents. In: Proceedings of the IEEE tenth workshop on multimedia signal processing, pp 393–398
Nitz K, Cruz W, Aradhye H, Shaham T, Myers G (2003) An image-based mail facing and orientation system for enhanced postal automation. In: Proceedings of 7th international conference on document analysis and recognition, pp 694–698
Norouzi A, Saba T, Rahim MSM, Rehman A (2012) Visualization and segmentation of 3D bone from CT images. Int J Acad Res 4(2):201–207
Oh S, Suen CY (1998) Distance features for neural network-based recognition of handwritten characters. Int J Doc Anal Recognit 1(1):73–88
Otsu N (1979) A threshold selection method from gray level histograms. IEEE Trans Syst Man Cybern 9(1):63–66
Pal U, Sinha S, Chaudhuri BB (2003) Multi-script line identification from Indian documents. In: Proceedings of the seventh international conference on document analysis and recognition, vol 2, pp 880–884
Phetchanchai C, Selamat A, Rehman A, Saba T (2010) Index financial time series based on Zigzag-perceptually important points. J Comput Sci 6(12):1389–1395
Rahim MSM, Rehman A, Faizal-Ab-Jabal M, Saba T (2011) Close spanning tree approach for error detection and correction for 2D CAD drawing. Int J Acad Res 3(4):525–535
Rahim MSM, Rehman A, Kumoi R, Abdullah N, Saba T (2012a) FiLeDI framework for measuring fish length from digital images. Int J Phys Sci 7(4): 607–618. doi:10.5897/IJPS11.1581
Rahim MSM, Rehman A, Sholihah N, Kurniawan F, Saba T, Mohamad D (2012b) Region-based features extraction in ear biometrics. Int J Acad Res 4(1):37–42
Raju SS, Pati PB, Ramakrishnan AG (2004) Gabor filter based block energy analysis for text extraction from digital document images. In: Proceedings of the first international workshop on document image analysis, pp 233–243
Rehman A (2010) Offline cursive character recognition based on heuristics techniques. PhD thesis, Universiti Teknologi Malaysia, pp 80–85
Rehman A, Mohamad D (2008) A simple segmentation approach for unconstrained cursive handwritten words in conjunction of neural network. Int J Image Process 2(3):29–35
Rehman A, Saba T (2011a) Document skew estimation and correction: analysis of techniques, common problems and possible solutions. Appl Artif Intell 25(9):769–787
Rehman A, Saba T (2011b) Performance analysis of segmentation approach for cursive handwritten word recognition on benchmark database. Digit Signal Process 21:486–490
Rehman A, Saba T (2012a) Features extraction for soccer video semantic analysis: current achievements and remaining issues. Artif Intell Rev. doi:10.1007/s10462-012-9319-1
Rehman A, Saba T (2012b) Analysis of advanced image processing to clinical and preclinical decision making with prospectus of quantitative imaging biomarkers. Artif Intell Rev. doi:10.1007/s10462-012-9335-1
Rehman A, Saba T (2013) An improved intelligent model for visual scene analysis and compression. Int Arab J Inf Technol IAJIT (ISI indexed) (accepted)
Rehman A, Kurniawan F, Mohamad D (2008a) Off-line cursive handwriting segmentation, a heuristic rule-based approach. J Inst Math Comput Sci (Computer Science Series) 19(2):135–139
Rehman A, Kurniawan F, Mohamed D (2008b) Off-line cursive character recognition based on hybrid statistical features. In: International graduate conference on engineering and science, UTM Skudai (IGCES, 08)

Rehman A, Mohamad D, Kurniawan F (2008c) Line and skew removal from off-line cursive handwritten words. Int J Res Sci 24(2):28–33
Rehman A, Mohamad D, Sulong G, Saba T (2009a) Simple and effective techniques for core zone detection and slant correction in script recognition. In: The IEEE international conference on signal and image processing applications (ICSIPA'09), pp 15–20
Rehman A, Mohamad D, Sulong G (2009b) Implicit vs explicit based script segmentation and recognition: a performance comparison on benchmark database. Int J Open Probl Comput Sci Math 2(3):352–364
Rehman A, Kurniawan F, Mohamad D (2009c) Neuro-heuristic approach for segmenting cursive handwritten words. Int J Inf Process IJIP 3(2):37–46. ISSN 0973-8215
Rehman A, Mohamad D, Kurniawan F (2009d) An automated approach to remove line from text bypassing restoration stage. In: Proceedings of 2nd IEEE international conference on computer, control and communication, pp 1–4
Rehman A, Saba T, Sulong G (2010) An intelligent approach to image denoising. J Theor Appl Inf Technol 17(1):32–36
Rehman A, Kurniawan F, Saba T (2011) An automatic approach for line detection and removal without characters smash-up. Imaging Sci J 59:171–182
Saba T (2012a) Offline cursive touched script non-linear segmentation. PhD thesis, Universiti Teknologi Malaysia, pp 102–115
Saba T (2012b) Offline cursive touched script non-linear segmentation, PhD thesis, Universiti Teknologi Malaysia, pp 133–138
Saba T, Rehman A (2011) Off-line cursive script recognition: current advances, comparisons and remaining problems. Artif Intell Rev 37(4):261–268. doi:10.1007/s10462-011-9229-7
Saba T, Rehman A (2012) Effects of artificially intelligent tools on pattern recognition. Int J Mach Learn Cybern. doi:10.1007/s13042-012-0082-z
Saba T, Rehman A, Sulong G (2009) ITS: using A.I. to improve character recognition of students with intellectual disabilities. In: International conference on software engineering and computer systems, UMP Malaysia, pp 5–9
Saba T, Rehman A, Sulong G (2010a) Improved offline connected script recognition based on hybrid strategy. Int J Eng Sci Technol 2(6):1603–1611
Saba T, Rehman A, Sulong G (2010b) Non-linear segmentation of touched Roman characters based on genetic algorithm. Int J Comput Sci Eng 2(6):2167–2172
Saba T, Rehman A, Sulong G (2011a) Cursive script segmentation with neural confidence. Int J Innov Comput Inf Control IJICIC 7(7):1–10
Saba T, Rehman A, Sulong G (2011b) Improved statistical features for cursive character recognition. Int J Innov Comput Inf Control IJICIC 7(9):5211–5224
Saba T, Sulong G, Rehman A (2011c) Document image analysis: issues, comparison of methods and remaining problems. Artif Intell Rev 35(2):101–118. doi:10.1007/s10462-010-9186-6
Saba T, Rehman A, Elarbi-Boudihir M (2011d) Methods and strategies on off-line cursive touched characters segmentation: a directional review. Artif Intell Rev. doi:10.1007/s10462-011-9271-5
Serra J (1982) Image analysis and mathematical morphology. Academic Press, London
Shridhar M, Badreldin A (1984) High accuracy character recognition using Fourier and topological descriptors. Pattern Recognit 17:515–524
Shridhar M, Kimura F (1995) Handwritten address interpretation using word recognition with and without lexicon. In: Proceedings of the IEEE international conference on systems, man and cybernetics, Piscataway, NJ, USA, vol 3, 2341–2346
Singh S, Hewitt M (2000) Cursive digit and character recognition on CEDAR database. In: Proceedings of international conference on pattern recognition, pp 569–572
Spitz AL (1997) Determination of the script and language content of document images. IEEE Trans Pattern Anal Mach Intell 19(3):235–245
Srihari SN, Shim YC, Ramanprasad V (1994) A system to read names and address on tax forms, Technical Report CEDARTR-94-2, CEDAR, SUNY, Buffalo, NY
Suen CY (1986) Character recognition by computer and applications. In: Young TY, Fu K-S (eds) Handbook of pattern recognition and image processing. Academic Press Inc., San Diego, CA, pp 569–586
Sulong G, Saba T, Rehman A (2010) Dynamic programming based hybrid strategy for offline cursive script recognition. In: IEEE second international conference on computer and engineering, vol 2, pp 580–584
Sulong G, Saba T, Rehman A, Saparudin (2009) A new scars removal technique of fingerprint images. In: IEEE international conference on instrumentation communication, information technology and biomedical engineering (ICICI-BME), Bandung, Indonesia, pp 31–35
Trier OD, Jain AK (1995) Evaluation of binarization methods for document images. IEEE Trans Pattern Anal Mach Intell 17:312–315

Uchida S, Taira E, Sakoe H (2001) Non-uniform slant correction using dynamic programming. In: Proceedings of 6th international conference on document analysis and recognition, vol 1, pp 434–438
Vamvakas G, Gatos B, Pratikakis I, Stamatopoulos N, Roniotis A, Perantonis SJ (2007) Hybrid off-line OCR for isolated handwritten greek characters. In: Proceedings of fourth IASTED international conference on signal processing, pattern recognition and applications, pp 197–202
Verma B (2002) A contour character extraction approach in conjunction with a neural confidence fusion technique for the segmentation of handwriting recognition. In: Proceeding of the 9th international conference on neural information processing, vol 5, pp 2459–2463
Verma B, Blumenstein M, Ghosh M (2004) A novel approach for structural feature extraction: contour vs. direction. Pattern Recognit Lett 25(9):975–988
Vijaya PA, Padma MC (2009) Text line identification from a multilingual document. In: Proceedings of the international conference on digital image processing, pp 302–305
Vinciarelli A, Luettin J (2001) A new normalization technique for cursive handwritten words. Pattern Recognit Lett 22:1043–1050
Wang D, Srihari SN (1991) Analysis of form images. In: Proceedings of first international conference on document analysis and recognition, Saint Malo, France, pp 181–191
Xingyuan L, Wen G (1999) A robust method for unknown structure form analysis. J Softw 10(11):1216–1224
Xiuling H, Yang Y, Zengzhao C, Ying Y, Cailin D (2006) Field extraction based on two-level regulated HMT in auto form processing. In: Proceedings of sixth international conference on intelligent systems design and applications, (ISDA'06), vol 2, pp 716–719
Yamada H, Nakano Y (1996) Cursive handwritten word recognition using multiple segmentation determined by contour analysis. IEICE Trans Inf Syst E79-D:464–470
Yong JY, Kim MK, Bana SW, Kwon YB (1997) Line removal and restoration of handwritten characters on the form documents. In: Proceedings of the fourth international conference on document analysis and recognition, vol 1, pp 128–131
Yoo J-Y, Kim M-K, Han SY, Kwon Y-B (1997) Line removal and restoration of handwritten characters on the form documents. In: Proceedings of the fourth international conference on document analysis and recognition, vol 1, pp 128–131
Yu B, Jain AK (1996) A generic system for form dropout. IEEE Trans Pattern Anal Mach Intell 18(11): 1127–1131
Yuan J, Tang Y, Suen CY (1995) Four directional adjacency graphs (FDAG) and their application in locating fields in forms. In: Proceeding of 3rd international conference on document analysis and recognition, Montreal, Canada
Zheng Y, Li H, Doermann D (2004) Machine printed text and handwriting identification in noisy document images. IEEE Trans Pattern Anal Mach Intell 26(3):337–353
Zheng Y, Liu C, Ding X (2002) Single character type identification. In: Proceedings of SPIE conference on document recognition and retrieval, pp 49–56
Zhou L, Lu Y, Tan CL (2006) Bangla/English script identification based on analysis of connected component profiles. In: Proceedings of 7th document analysis systems, pp 243–254