ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

INTELLIGENT CHARACTER RECOGNITION (ICR): A Novel Algorithm to Extract Text from a Scanned Form Based Image

Kranthi Kumar.K[1], Madhuri Venkata Saroja Muvvala[2], Pasikanti Susruthi Divya Sruthi[3], Pilla Dinesh[4]
[1]Assistant Professor, Department of IT, SNIST, Yamnampet, Ghatkesar, Hyderabad, AP, INDIA
[2-4]M.Tech (CN&IS), Department of IT, SNIST, Yamnampet, Hyderabad, AP, INDIA.

Abstract: Image Processing is nowadays considered to be a favourite topic in the IT industry. It is a field under Digital Signal Processing. One of its major applications is Intelligent Character Recognition (ICR).Intelligent character recognition, usually abbreviated to ICR, is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. ICR enables the computer or machine to visualize an image and extract text from it such that it can edit the text, store it, display or print without any scanning and apply techniques like text to speech and text mining to it.
In this paper, we propose a novel algorithm to extract text/characters from a scanned form image. Our system consists of various stages like 1) uploading a scanned image from machine/computer 2) Extraction of text zone from the image 3) recognition of the text and 4) applying post processing techniques(error correction and detection methods). In addition, discussed the form image registration technique, image masking and image improvement techniques are implemented in our system as part of the character image extraction process. In our experiment we show that, the proposed system will get good results then the existing systems and trying to improve the efficiency and accuracy of recognizing the characters from a scanned form image.

Keywords: Form-based ICR, Character Image Extraction, Image Registration, Image Masking and Image Extraction.

INTRODUCTION

Digital Image Processing is a rapidly evolving field with the growing applications in science & engineering. Image Processing holds the possibility of developing an ultimate machine that could perform visual functions of all living beings. The term Digital Image Processing generally refers to the processing of a two-dimensional picture by a digital computer i.e. altering an existing image in the desired manner. Since the image processing is a visual task, the foremost step is to obtain an image. An image is basically a pattern of pixels (picture elements) thus a digital image is an array of real & complex numbers represented by finite number of bits.

Manual data entry from hand-printed forms is very time consuming - more so in offices that have to deal with very high volumes of application forms (running into several thousands). A form based Intelligent Character Recognition (ICR) System has the potential of improving efficiency in these offices using state-of-the-art technology. An ICR system typically consists of several sequential tasks or functional components, viz. form designing, form distribution, form registration, field-image extraction, feature-extraction from the field-image, field-recognition (here by field we mean the handwritten entries in the form). At the Centre for Artificial Intelligence and Robotics (CAIR), systematic design and development of methods for the various sub-tasks has culminated into complete software for ICR. The CAIR ICR system uses the NIST (National Institute for Standards and Technology, USA) neural networks for recognition [1,2,3]. For all the other tasks such as form designing, form registration, field-image extraction etc. algorithms have been specially designed and implemented. The NIST neural networks have been trained on NIST's Special Database 19 [4,5,6]. The classification performance is good provided the field, i.e. the handwritten character entry in the form, is accurately extracted and appropriately presented to the neural network classifiers. Good preprocessing techniques preceding the classification process can greatly enhance recognition accuracy [7,8]. It is always fascinating to be able to find ways of enabling a computer to mimic human functions, like the ability to read, to write, to see things, and so on. ICR enables the computer or machine to visualize a image and extract text from it such that it can edit the text, store it, display or print without any scanning &apply techniques like text to speech and text mining to it.

When the object to be matched is presented then our brains or in general recognition system starts extracting the important features of the object that includes color, depth, shape & size. These features are stored in the part of the memory. Now the brain starts finding the closest match for these extracted features in the whole collection of objects, which is already stored in it. This we can refer as standard library. When it finds the match then it gives the matched object or signal from the standard library as the final result. For humans character recognition seems to be simple task but to make a computer analyze and finally correctly recognize a character is a difficult task. ICR is one such technique that gives the power of vision to the computer to extract data from images and find important data from it and make it computer editable.

In this paper, we propose a novel algorithm to extract text from a scanned form-based image, which is used to extract text information from the scanned copy of form and also discussed about Image registration, Image masking and Improvement techniques. We concentrated on Skewing methodologies also. In next coming sections are articulated as related work, proposed system, conclusion and references.

RELATED WORK

A digitized image is, after all, just a collection of numbers. For binary images, every point or pixel is assigned a value of either 0 or 1, for gray level images pixel values range from 0 to 255, and for color images pixel values usually consist of three numbers, each in the range of 0 to 255. While the ensuing discussion is valid for any type of image, for the sake of simplicity, only binary images will be addressed.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

Recognition technologies may be classified as

1. Statistical
2. Semantic and
3. Hybrid.

In the following, these methodologies are discussed. Only handwritten numeric characters or digits will be considered, such that the character recognition algorithms return only the values 0, 1, 2...9 or "reject".

Statistical Approach

Since every electronic image of a digit consists of pixel values that are represented by a spatial configuration of "0"s and "1"s, a statistical approach to image character recognition would suggest that one look for a typical spatial distribution of the pixel values that characterize each digit. In general, one is searching for the statistical characteristics of various digits. These characteristics could be very simple, like the ratio of black pixels to white pixels, or more complex, like higher order statistical parameters such as the third moments of the image.

The general flow of statistic based character recognition algorithms is as follows:

1. Compute the relevant statistics for a digitized image
2. Compare the statistics to those from a predefined database.

In general, most statistical methods of character recognition work well for digits that do not vary much from an "ideal" or predefined digit. Unfortunately, in reality, handwritten images demonstrate a large variance. Thus, some additional approaches are required to solve the character recognition problem.

Semantic Approach

Digitized images of handwritten characters indeed consist of pixels. However, a fact that most statistical methods ignore is that the pixels also form lines and contours. This is the essential point of the semantic approaches to character recognition: first recognize the way in which the contours of the digits are reflected in the pixels that represent them and then try to find typical characteristics or relationships for each digit.

The steps of a semantic based classifier for character recognition are as follows:

1. Find the starting point of a contour.
2. Start tracing the contour.
3. Identify the characteristics of the contour while tracing it: "up", "down", "diagonal up", "arc", "loop", etc.
4. Search the database for a description similar to the one obtained. Technically, this would be executed by representing the descriptions as a logic tree (graph) and then by matching the graph against the graphs contained in the database.

The main problem with semantic methods is their reliance on correct extraction of character contours from the electronic images. For instance, if a character image is broken, then a semantic method may fail to trace the character's contour correctly. On the other hand, a statistical approach could still reflect broken character image statistics with sufficient accuracy to enable correct identification.

Hybrid Approach

It is clear that statistical and semantic approaches to character recognition have specific advantages and disadvantages. The obvious question: Is it possible to combine the best of both methods? The answer: To a certain extent, yes, it is possible to develop algorithms that are part statistical and part semantic in an effort to leverage the advantages of both. Top Image System's proprietary T.i.S. ICR/OCR reflects such a hybrid approach and, in many cases, overcomes the problems associated with the statistical and semantic methods when utilized independently.

There is another step which can be taken, given extant technologies and methodologies, to obtain the best of all available recognition algorithms. Today, there are substantial number of good ICR recognition engines available. Each of these engines has its own specific strengths and weakness. Each engine, on a particular type of image or document, performs better than its peers, on another, worse. Recognizing this flaw common to all ICR engines, T.i.S. analyses the relative strengths and weaknesses of the different recognition engines. This knowledge allows for the creation of unique "voting" algorithms which draw on the strengths of various engines optimizing recognition results.

Character extraction problems have been solved by some researchers using variant methods. Neves [9] proposed the cell extraction method for Table Form Segmentation which consists of steps such as initially locating and extracting the intersections of table lines. The weakness of this method is that the process involved complicated table extractions. Chen and Lee [10] presented a novel approach using a gravitation-based algorithm. However, in their work, some field data could not be extracted correctly, which led to mis-extraction. Tseng and Chuang proposed the stroke extraction method [11] and then used it for the Chinese characters.However, the method did not solve the overlapping problems. Liolios et al[12] described a system for form identification based on power spectral density of the horizontal projection of the blank form to obtain the feature vectors. Here too, the overlapping problem has not been addressed.

Lines are the essential elements of a form and their extraction is an important process before data extraction. This is evident in the works of researchers such as [13,14, 15,16,] However, in these researches, data extraction is degraded when characters are overlapped by boxes' lines. Boatto et al. [17] proposed an interpretation system for land register maps. Their system could only work on the assumption that the symbols consist of certain topological structure. On the other hand, Wang and Srihari [18] proposed a system to analyze form images. Their shortcoming was that the system could not patch the broken characters well in all conditions; it would cause the data to further deteriorate in some cases. [19,20] proposed Intelligent Form Processing systems (IFP). Similarly, the setback here was that the processes of overlapped characters in related works usually lose partial information of the character and the broken strokes are not well reconstructed before recognition, such that the performance of character recognition is degraded. In this study, we propose simple algorithms based on the differentiating characteristics of

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

the shape of the boxes' lines and the handwritten characters lines.

PROPOSED SYSTEM

Fig.1. shows the general scheme of the proposed system to extract text/characters from a scanned form image. The following sections are discussed about various steps involved in the proposed system

Binarization

It converts an image from colour or greyscale to black-and-white (called a "binary image" because there are two colours). It converts the acquired form images to binary format, in which the foreground contains logo, the form frame lines, the pre-printed entities, and the filled in data.

(a) **Global Gray Thresholding**

Global Thresholding algorithms use a single threshold for the entire image. A pixel having a gray level lower than the threshold value is labelled as print (black), otherwise background (white).

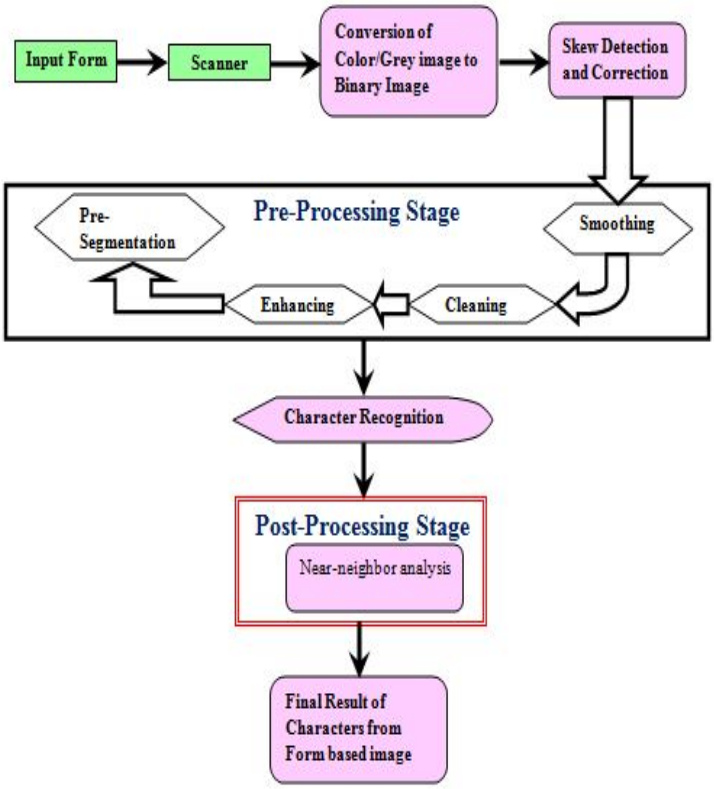

Fig 1: Architecture of the Proposed System

(b) **Local Gray Thresholding**

Local Thresholding algorithms compute a designated threshold for each pixel based on a neighbourhood of the pixel. A pixel having a gray level lower than 14 the threshold value is labelled as print (black), otherwise background (white). It is used for slowly varying background. Some of the methods applied for this are Bersen method, Niblack method, etc.

(c) **Feature Thresholding**

Feature Thresholding is just like local gray Thresholding but is applied for abruptly changing background.

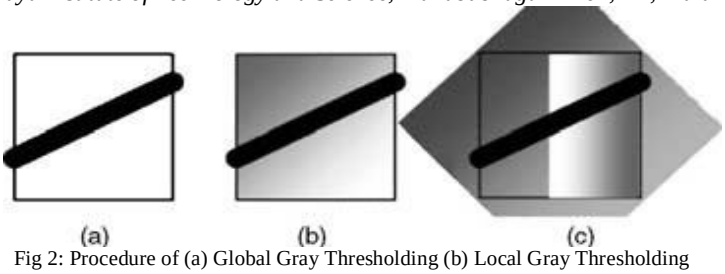

Fig 2: Procedure of (a) Global Gray Thresholding (b) Local Gray Thresholding (c) Feature Thresholding

2. Scanning: Form Identification and Image Registration

A form identification stage matches the extracted features (e.g., line crossings) from an incoming form against those extracted from each modelled design in order to select the appropriate model for subsequent system processes. Form registration is done to align the object so that it can be processed. Fig 3 shows the scanned copy which we are considered in our experiment.

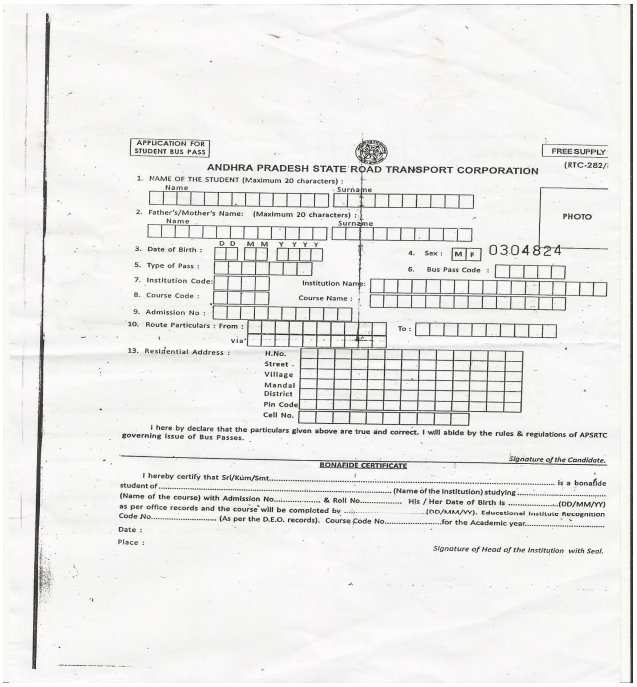

Fig 3: scanned form of APSRTC bus pass application

Algorithm: Form Registration

Input : Scanned form image.

Output: Form image after skew and shift correction.

begin

1. *Extract adequate portions from the top and bottom half of the form image for detecting the two horizontal sides of the rectangle. Similarly for detecting the vertical sides, extract sub-images from the left and right halves of the form image.*
2. *Divide the sub-images into strips and project each strip. Using the projection values along the scan lines detects the line segments in each strip and then the corresponding start points.*
3. *Use the line tracing algorithm similar to that in with a 3 × 3 window for connectivity checking. Having obtained segment*

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

points using line tracing, fit a line to these points using the pseudo-inverse method to obtain the slope.

4. *Starting with the middle strip in each sub-image, merge the line segments with the line segments in the strips on either directions of the middle strip to obtain full length lines. This results in four sets of lines corresponding to the four sub-images called the top line set, the bottom line set, the left line set and the right line set.*
5. *Identify a pair of lines, one from the top line set and the other from the bottom line set as the top edge and the bottom edge of the rectangle respectively, if they are almost parallel to each other and the perpendicular distance between them is equal to the height of the rectangle. Similarly identify the left and right edges of the bounding rectangle using the left line set and the right line set.*
6. *To improve the estimates, discard the outliers from the coordinates array of points of the detected edges. Fit a line using least squares for points in the new array and return this array along with the estimated slope and offset. Use the slope values of the four lines to assess the skew in the form and rotate the form for correcting this skew. Perform the same transformation on the edge points and recompute the slope and offset values of the edges after rotation. Use these new offset values for shift correction.*

end

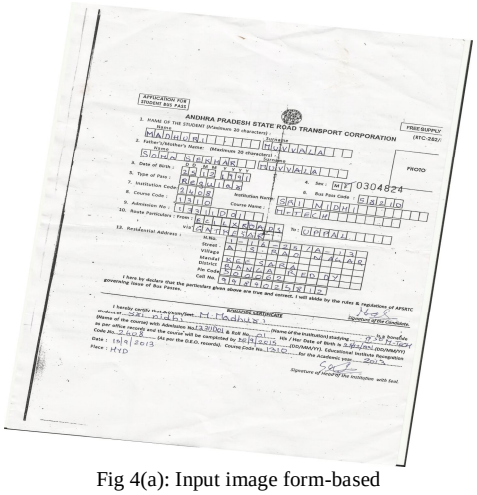

Fig 4(a): Input image form-based

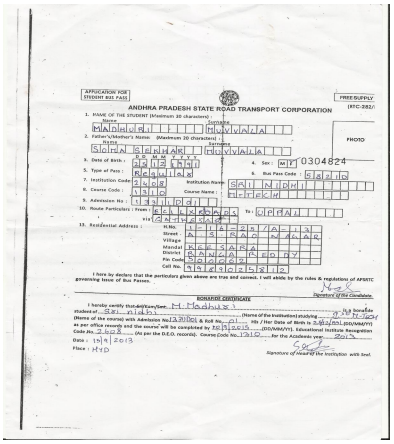

Fig 4(b): Form-based image after skew and shift correction

Table 1: Actual vs. Measured skew angles by the Registration Algorithm

Actual Skew	Measured Skew
2^0	2.042^0
3^0	2.982^0
6^0	5.996^0
8^0	7.843^0
13^0	12.943^0
18^0	17.94^0

3. Skew Detection and Correction

3.1. Local Registration and Field Box Masking

A robust and efficient algorithm for registration of scanned images was presented. The form registration algorithm performs a global registration of the form and though an essential ingredient of the processing scheme is not sufficient. Form printing and subsequent form scanning for ICR introduces non-uniform distortions in the form image. This necessitates a local registration of field boxes after a global registration of the form image has been performed. This is because the skew and shift parameters computed by the registration algorithm for the current scanned form are used to correct the box position values stored by the system during the form design process. However these corrected boxes need not exactly coincide with the boxes in the registered image of the scanned form. In Figure 4, the light gray boxes represent the box positions corrected for the computed skew and shift and dark gray boxes represent the actual box positions in the registered image. Moreover the field boxes themselves may undergo a structural change due to the process of printing and scanning further complicating the process of field box masking.

Accurate field box masking is essential for character image extraction because if portions of the box remain in the final character image presented to the neural network classifier the results are often inaccurate. To circumvent these problems the following correlation based algorithm has been devised to perform local registration and masking of the field boxes.

Algorithm: Mask form image field boxes.

Input : Form image, list of ideal registration points for field boxes (as recorded by the form design module.)

Output: Masked image of the form, list of registration points corrected for the current form.

Begin

1. *Initialize corrected registration points list as an empty list.*
2. *For every field box perform steps 3 to 8.*
3. *Set Top Left (x, y) to top left corner coordinate of current field box from the list of ideal registration*
4. *points.*
5. *Set Bottom right (x, y) to bottom right corner coordinate of current field box from the list of ideal registration points.*
6. *In the neighborhood of the field box ideal position, locate the position with maximum correlation in terms of the number of matching ink points. (The neighborhood is defined by an N x N grid centered at the ideal field box position.) Label this box as Max corr box and its corner points as Max corr*

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

top left and Max corr bottom right respectively, and the correlation value as Max corr.

7. *Stretch each side of the Max corr box, one at a time, to a maximum distance of DELTA and store the resulting corner coordinates each time the correlation exceeds THRESHOLD (= THRESH PERCENT * Max corr).*
8. *Draw on the form image, in background colour, the Max corr box, and each box whose coordinates have been stored in step 6.*
9. *Append Max corr box corner coordinates to corrected registration points list.*
10. *Return the masked image and the corrected registration points list.*

End

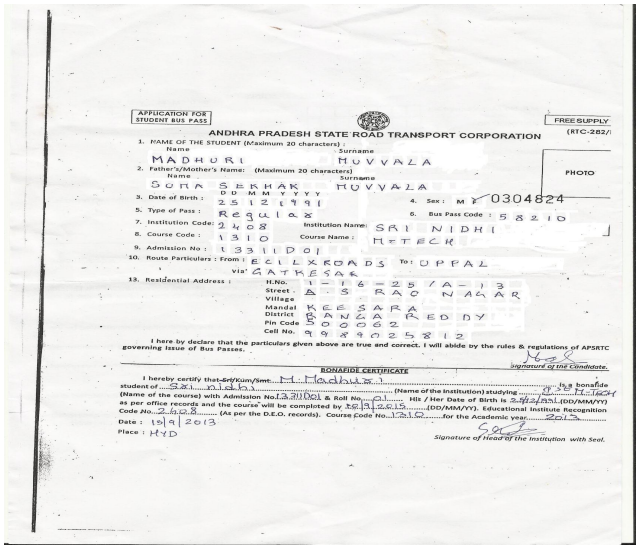

Fig 5: Scanned image after masking

3.2 De-skew

If the document was not aligned properly when scanned, it may ne be tilted a few degrees clockwise or counter clockwise in order to r lines of text perfectly horizontal or vertical. During the document scanning process, the whole document or a portion of it can be fed through the loose-leaf page scanner. Some pages may not be fed straight into the scanner, however, causing skewing of the bitmapped images of these pages. So, document skew often occurs during document scanning or copying Fig 6(a) shows it. This effect visually appears as a slope of the text lines with respect to the *x*-axis, and it mainly concerns the orientation of the text lines.

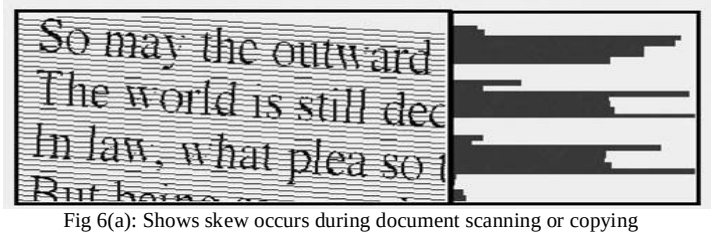

Fig 6(a): Shows skew occurs during document scanning or copying

By applying the principles of trigonometry, the images can be corrected fig 6(b).

Fig 6(b): Document view after De-Skew

3.3 Slant Detection

The character inclination that is normally found in cursive writing is called slant. Figure below shows some samples of slanted handwritten numeral string. Slant correction is an important step in the pre-processing stage of both handwritten words and numeral strings recognition. The general purpose of slant correction is to reduce the variation of the script and specifically to improve the quality of the segmentation candidates of the words or numerals in a string, which in turn can yield higher recognition accuracy. Fig 7(a) and Fig 7(b) shows the images before and after slant detection.

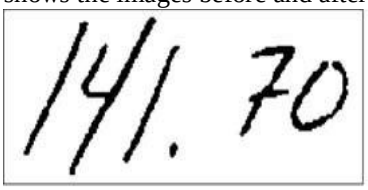

Fig 7(a):Slanted digits

Fig 7(b)Corrected digits

3.4 Image Extraction

It extracts pertinent data from respective fields and preprocesses the data to remove noise and enhance the data. This step extracts the field in which characters to be recognized is present after the layout analysis.

4. Pre-processing Stage

OCR software often "pre-processes" images to improve the chances of successful recognition. Techniques include: Smoothing, Cleaning, Enhancing and Pre-Segmentation

4.1 Smoothing

Smoothing operations in gray level document images are used for blurring and for noise reduction. Blurring is used in pre-processing steps such as removal of small details from an image. In binary (black and white) document images, smoothing operations are used to reduce the noise or to straighten the edges of the characters, for example, to fill the small gaps or to remove the small bumps in the edges (contours) of the characters. Smoothing and noise removal can be done by filtering. Filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. There are two types of filtering approaches: linear and nonlinear.

These masks are passed over the entire image to smooth it, and this process can be repeated until there is no change in the image. The pixel in the centre of the mask is the target. Pixels overlaid by any square marked “X” are ignored. If the pixels overlaid by the squares marked “=” all have the same value, that is, all zeros, or all ones, then the target pixel is forced to match them to have the same value, otherwise it is not hanged. These

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

masks can fill or remove single pixel indentation in all the edges, or single bumps.

Fig 8: a) Before smoothing b) after smoothing

4.2 Cleaning

It is done to separate and remove the preprinted entities while preserving the user entered data (filled in data) by means of image-processing techniques (base line removal). The form outline information from the knowledge base is utilized for cleaning procedure.

4.3 Enhancing

It reconstructs strokes that have been removed during cleaning. If any filled in data is written over baseline, then during cleaning it is also removed leading to discontinuity in the character which may hamper its recognition.

4.4 Pre-Segmentation

It separates the characters(as shown in Fig 9) from pre-printed texts like in bank form, the preprinted texts are 'signature', 'amount', etc.. After the pre-segmentation process, we have a clear view of the characters to be recognized. The following diagram is the procedure layout of preprocessing stage

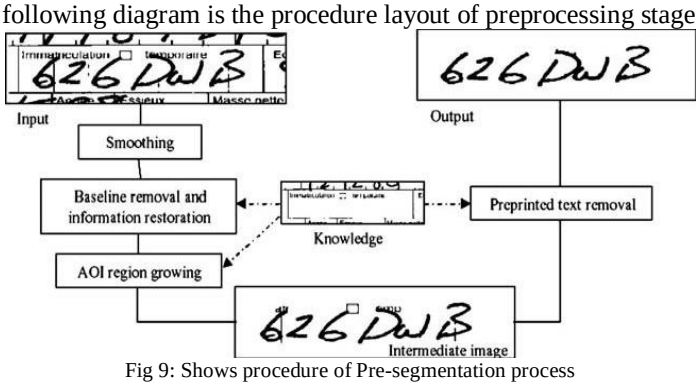

Fig 9: Shows procedure of Pre-segmentation process

5. Character Recognition

Recognition is to map the given pattern with internally stored database. In other words recognition is a process of matching the visible or the presented signal with the standard signal. When a signal or any image is input to the recognition system it extracts the important features like the size, depth, shape, color. These features are stored in the part of memory. After extracting the information, the recognition system finds the closest match of the input signal or the image with the standard library of signals or images. The principle behind the proposed method of recognition is that the basic geometry of a pattern or character is retained even after deformations. If these geometrical features can be extracted from the input character then the character can be matched with the standard characters in the library.

Recognition is done in the following steps:-

- Separation of characters
- Normalization.
- Thinning.
- Singular point determination.
- Grid formation.
- Line Detection.
- Character recognition.

5.1 Separation of Characters

The basic idea behind the separation of character is to search the blank spaces between the individual characters. When the selected word is scanned, right from the first character then each time it will come across a pixel i.e. tiny dots which indicate the presence of a part of character. This will continue till it gets the two or three continuous blanks, which is the indication of the end of a single character. In this way the characters can be separated. The individual characters are then given to the next state as input, which is basically a normalization step.

5.2 Normalization of Characters

The characters written by hand vary greatly in size and shape. To account for variety in shape, the character is normalized. By normalization, we mean that the character is made to fit into a standard size square. This is accomplished by filling a 2-D array by '1's or '0's depending upon the presence or absence of text pixel in that particular area. This size of array is chosen by trial and error method & the value that gives the best results is fixed. Characters of any size and shape can be processed and matched with the normalization technique.

Let the height & width of the character is 'h' & 'w' respectively. Thus normalizing factor x normal and y normal in X & Y directions are given by

- x_{normal} = array size / w
- y_{normal} = array size / h

Then if the pixel on the screen has the co-ordinates x' and y' relative to the top left corner of the bounding rectangle of the character, the X and Y coordinate of that pixel in the standard size array are –

- $x = x' * x_{normal}$
- $y = y' * y_{normal}$

Thus the final result is the 2-D integer array i.e. the standard approximation of the character. If the color of the pixel on the screen is not the same as background color then the corresponding array element is filled as '1' else it is filled as '0'. The normalization can be briefly represented by the following fig 10.

Fig 10: Normalization of character 'A'

5.3 Thinning

Thinning algorithm is applied to all normalized characters. When the character is normalized, some redundant 1's are also stored in the standard array. This affects the parameters of the character i.e. thickness. Thinning reduces the thickness of the characters to their actual skeleton as well as removes all the redundant ones with the help of eight standard deletion masks i.e. two for each direction (top, bottom, left & right). The lines forming the characters reduce to a width of one pixel.

However by deletion of the extreme pixels, some important pixels may also be deleted that may result the discontinuity of the image. Hence four non deletion masks are used. If any of the four masks is satisfied then the picture is retained. Thinning also

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

helps in detection of Singular Points & lines. The skeleton obtained must have the following properties:

- Must be as thin as possible
- Connected
- Centered

Thinning process is explained in the below Fig 11.

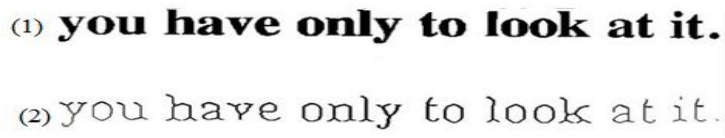

Fig 11: The above line (1) is extracted before thinning while (2) is after thinning.

5.4 Grid Formation

In any character there are certain geometrical similarities. These properties are local to the particular area. For example if we consider the alphabet 'A', the topmost point is a vertex with two lines. Similarly a triangle is formed by three lines. However when we consider the hand written characters it is possible that these properties might be slightly displaced. But the properties will occur close to the exact positions. By grid formation we aim to collect information about these properties & their relative positions.

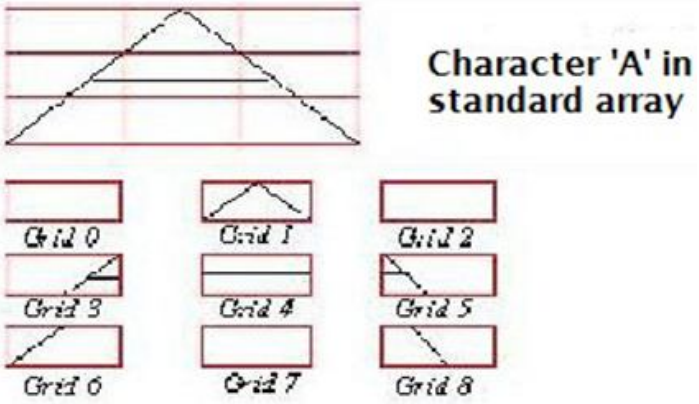

Fig 12: Shows the formation of grid

In grid formation the standard array is divided into nine squares or grids as shown in Fig 11. These grids are then analyzed to find out the number of lines & SPs. Each grid will have a set of lines & or singular points. If none of these characteristics are present then the variables for that particular grid will be set to 0. Thus its lines & SPs will characterize each grid.

5.5 Singular Point Determination:-

Singular point can be defined as any point with a degree greater that two. Thus, SP's are those points where branching occurs. The degree of a point is equal to the number of lines emerging from that point. The following figure shows singular points for character 'A' –

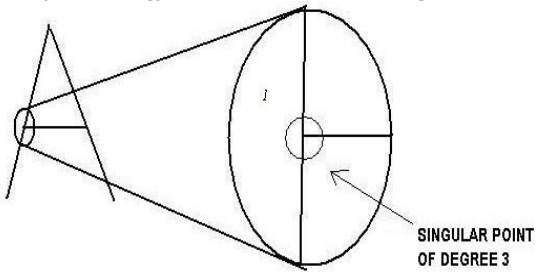

Fig 13(a): Singular Point Detection

If the point is on line, the degree is 2 & if the point is end point of a line, then its degree is 1. However for a point where branching occurs the degree is 3 or greater. These SP's are important characteristics in a particular character, and their information can be used to differentiate between different characters. To determine whether the point is SP or not, it is necessary to find out the total number of lines originating from the point. To do so we check the pixels in 5 by 5 neighborhoods as shown in the figure below-

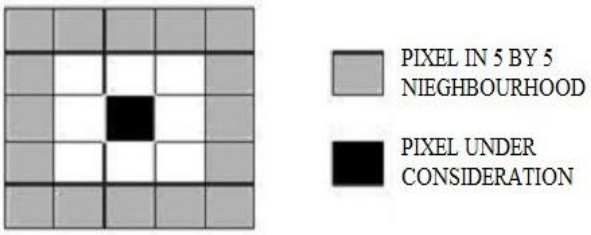

Fig 13(b): shows 5X5 Neighbourhood consideration

Here we check for the total number of 0 to 1 transitions. If the 0 to 1 transitions are more than 2, it indicates that the number of lines from that point are also greater that two. Hence any point which has more than two 0 to 1 transitions is taken as SP. Once the co-ordinates of SP are known, the SP is assigned to a grid.

5.6 Line Detection:-

Line detection is done on the basis of direction flags set for each pixel in the grid. The direction flags are first set for each pixel by analyzing the neighboring pixels of the current pixel. The directions are numbered from 1 to 8 as shown in the figure in the back page. To detect the lines the direction flags assigned to every pixel are counted & the direction that has majority of the pixels is assigned to the line. This method not only gives the number of lines in the grid but also direction of each line. If the grid doesn't have any pixel it is said to have no lines.

5.7 Character matching:-

The last step in OCR is the matching of the scanned input character with the standard one. The characters are represented as a set of nine sets of parameters, one set corresponding to a grid. Thus each grid will have its own set of parameters.

SET 1:

- Number of lines in that grid.
- Number of SPs in that grid.
- Direction of each of the lines.

SET 2:

- Nine grid structures of the form of set1.
- The actual character.
- Total number of SPs in the character.
- Total number of grids marked as important.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.6, Pages : 47- 55 (2013)
Special Issue of ICETEM 2013 - Held on 29-30 November, 2013 in Sree Visvesvaraya Institute of Technology and Science, Mahabubnagar – 204, AP, India

- Grid number of the important grids.

The above data is retrieved from the input character & it is matched with the inbuilt library of characters. All characters in the library also have above structure format. The matching process is done on the point basis i.e. whenever a characteristic of the input character matches with that of the library the points for that character are increased. The grids are very crucial in the matching. If the characteristics of the grids don't match the points of the character are made '0'. In the end the character scoring the maximum number of points is elected as the recognized character.

6. Post-processing

OCR accuracy can b e increased if the output is constrained by a lexicon – a list of words that are allowed to occur in a document. This might be, for example, all the words in the English language, or a more technical lexicon for a specific field. This technique can be problematic if the document contains words not in the lexicon, like proper nouns. Tesseract uses its dictionary to influence the character segmentation step, for improved accuracy.

The output stream may be a plain text stream or file of characters, but more sophisticated OCR systems can preserve the original layout of the page and produce, for example, an annotated pdf that includes both the original image of the page and a searchable textual representation. "Near-neighbour analysis" can make use of co-occurrence frequencies to correct errors, by noting that certain words are often seen together. For example, "Washington, D.C." is generally far more common in English than "Washington DOC". Knowledge of the grammar of the language being scanned can also help determine if a word is likely to be a verb or a noun, for example, allowing greater accuracy.

CONCLUSION

To conclude, a robust algorithm has been described for measuring and correcting the skew and shift values that are present in a scanned form image. Subsequently, three techniques, viz. (*i*) image registration, (*ii*) field ox masking and (*iii*) pre-processing techniques, that together comprise the handwritten character extraction process have been presented. The necessity and impact of these methods on the overall performance of the OCR system has been systematically illustrated by examples. The effectiveness of these algorithms has been convincingly proved by the fact that the system performed with adequate accuracy in real life recruitment exercises requiring the processing of handwritten application forms.

From all the above explanations, it is observed that though the character recognition is a very small part of a very vast field of Digital Signal Processing, it is considered to be a boon to the banking institutions for signature recognition. The other applications include pattern recognition of the scanned digital images from the satellite and comparing them with the previous images. By doing this, the prediction about the climate can be made effectively. This can be implemented with a proper accuracy giving rich dividends in branch of banking, satellite communication as well as in other institutions where such signature or any character recognition is necessary.

REFERENCE

[1] M. D. Garris, C. L.Wilson, and J. L. Blue. Neural network-based systems for handprint ocr applications. *IEEE Transactions on Image Processing*, 7(8):1097–1110, August 1998.

[2] P. J. Grother. Karhunen lo`eve feature extraction for neural handwritten character recognition. *Applications of Artificial Neural Networks III, SPIE, Orlando*, 1709:155–166, April 1992.

[3] P. J. Grother. Karhunen lo`eve feature extraction for neural handwritten character recognition. *NIST Internal Report 4824*, April 1992.

[4] Patrick J. Grother. Nist special database 19 handprinted forms and characters database. Technical report, NIST, March 1995.

[5] M. D. Garris, J. L. Blue, G. T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C. L. Wilson. Nist form-based handprint recognition system. *NIST Internal Report 5469 and CD-ROM*, July 1994.

[6] P. J. Grother. Handprinted forms and characters database, nist special database 19. *NIST Technical Report and CD-ROM*, March 1995.

[7] Z. Shi and V. Govindraju. Character image enhancement by selective region-growing. *Pattern Recognition Letters*, 17:523–527, 1996.

[8] G. Nagy. Twenty years of document image analysis in pami. *IEEE Trans. Pattern Anal. And Mach. Intell.*, 22(1):38–62, January 2000.

[9] Neves, L.A.P., J.M. De Carvalho, J. Facon and F. Bortolozzi, 2006. A new table extraction and recovery methodology with little use of previous knowledge. Proceedings of the International Workshop on Frontiers in Handwriting Recognition, (FHR'06), La Baule. France.

[10] Chen, J.L. and H.J. Lee, 1996. Field data extraction for form document processing using a gravitationbased algorithm, Pattern Recognit., 34: 1741-1750. DOI: 10.1016/S0031-3203(00)00115-1

[11] Tseng, L.Y. and C.T. Chuang, 1992. An Efficient Knowledge-based stroke extraction method for multifont Chinese character. Pattern Recognition, 25: 1455-1458. DOI: 10.1016/0031-3203(92)90119-4

[12] Liolios, N., N. Fakotakis and G. Kokkinakis, 2002. On the generalization of the form identification and skew detection problem. Pattern Recognit., 35: 253-264. DOI: 10.1016/S0031-3203(01)00030-9

[13] Pizano, A., 1992. Extracting line features from images of business forms and tables. Proceedings of the 11th IPAR International Conference on Pattern Recognition, Aug. 30-Sep. 3, IEEE Xplore Press, The Hague , Netherlands, pp: 399-403. DOI:10.1109/ICPR.1992.202008

[14] Mandal, S., S.P. Chowdhury, A.K. Das and B. Chanda 2005. A hierarchical method for automated identification and segmentation of forms. Proceedings of 8th International Conference on Document Analysis and Recognition, Aug.

29-Sep. 1, IEEE Xplore Press, India, pp: 705-709. DOI: 10.1109/ICDAR.2005.17

[15] Guillevic, D. and C.Y. Suen, 1993. Cursive script recognition: A fast reader scheme. Proceeding of 2nd International Conference on Document Analysis and Recognition, Oct. 20-22, IEEE Xplore Press, Tsukuba Science City, Japan, pp: 311-314. DOI: 10.1109/ICDAR.1993.395725

[16] Mandal, S., S.P. Chowdhury, A.K. Das and B. Chanda, 2006. Fully automated identification and segmentation of form document form processing, springer, Comput. Graphics, 32: 953-961. DOI: 10.1007/1-4020-4179-9_139

[17] Boatto, L., V. Consorti, M. De Buono, S. Di Zenzo and V. Eramo *et al*., 1992a. An interpretation system for land register maps. Computer, 25: 25-33. DOI:10.1109/2.144437

[18] Wang, D. and S.N. Srihari, 1994. Analysis of form images. Int. J. Pattern Recognit., 8: 1031-1031.

[19] Casey, R.G. and D.R. Ferguson, 1990. Intelligent forms processing. IBM Syst. J., 29: 435-450. DOI: 10.1147/sj.293.0435

[20] Casey, R., D. Ferguson, K. Mohiuddin and E. Walach, 1992. Intelligent forms processing system. Machine Vision Appl., 5: 143-155. DOI:10.1007/BF02626994

ABOUT AUTHORS

Kranthi Kumar K[1]. is currently with Sreenidhi Institute of Science and Technology, Hyderabad, India, working as Asst. Professor in the department of Information Technology. He has 11 years of experience in Teaching. Graduated in B.Tech (CSE) from JNTU Hyderabad in 2003. Masters Degree in M.Tech (CSE), from JNT University, Anantapur, AP in 2007. He is pursuing his Ph.D in Computer Science & Engineering from JNT University, Hyderabad. He has 15 publications in various national/international journals and conferences. His areas of Interests are Image Processing, Content Based Image Retrieval, Information Retrieval Systems, Database Management Systems, Distributed Databases, and Computer Networks.

Madhuri Venkata Saroja Muvvala[2], is studying M.Tech (CN&IS) in Sreenidhi Institute of Science and Technology, Yamnampet, Gatkesar, Hyderabad. She completed B.Tech in ECE in the year 2013 from Vignan Institute of Technology and Aeronautical Engineering, Deshmukhi, Hyderabad. Her Areas of Interest are Image Processing, Communication Systems and Computer Networks & Security.

Pasikanti Susruthi Divya Sruthi[3], is studying M.Tech (CN&IS) in Sreenidhi Institute of Science and Technology, Yamnampet, Gatkesar, Hyderabad. She completed B.Tech in ECE in the year 2013 from CVR College of Engineering, Mangalpally, Hyderabad. Her Areas of Interest are Image Processing, Communication Systems , Digital Signal Processing and Computer Networks & Security.

Pilla Dinesh[4], is studying M.Tech (CN&IS) in Sreenidhi Institute of Science and Technology, Yamnampet, Gatkesar, Hyderabad. He completed B.Tech in ECE in the year 2013 from Viswanadha Institute of Technology and Management (VITAM), Visakhapatnam. His Areas of Interest are Image Processing, Communication systems, Object Oriented Programming and Computer Networks.