A - Survey and Analysis of Segmentation, Feature Extraction and Classification in OCR System

KEYWORDS: Segmentation, Feature Extraction, Classification

Modi Gaurav S	Chauhan Nandish
ME (CSE Student), S. P. B. Patel Engineering College, Mehsana, Gujarat	CSE-IT Department, S. P. B. Patel Engineering College, Mehsana, Gujarat

ABSTRACT To use historical data in and across the world we have to make it searchable over the internet. But old historical data is printed on paper. A direct solution to this problem is the use of character recognition system to convert document images into text followed by the use of existing text search engines to make them available for the public at large via an internet, so the vision is to develop an OCR system which can convert the printed document in editable document. There are many languages and scripts in all over the world so, there is a need of OCR system for all the scripts which is in development process. In OCR, the system include the scanning of document, cleaning the noise, skew detection and correction text and non-text classification, text line detection and segmentation. This paper contains a basic survey and analysis of Segmentation, Feature Extraction and classification in OCR system.

I. INTRODUCTION

Optical Character Recognition Is the most important gift given by Computer science to the mankind. It has made lot off tedious work easy and speedy. Optical Character Recognition is the mechanical or electronic translation of scanned image of handwritten, typewritten or printed text word-keeping system in an office or to publish the text on a website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning artifacts, and apply technique such as machine translation, text-to-speech and text meaning to it.OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

The various stages of OCR techniques are: [1]

(1.) DIGITIZATION: In this phase the printed or hand written Data converted in to the digital form by scanning the document or by digitizer.

(2.)PREPROCESSING: This phase performs various operatio- ns like binarization, contour smoothing, noise reduction , skew detection and finally sketeletonization of the digital Image [1][2].

(3.) SEGMENTATION : Segmentation is an integral part of Any Text based recognition system. Segmentation phase Include basically three phases, i.e. line segmentation, character segmentation, word Segmentation [2].

(4.) FEATURE EXTRACTION AND CLASIFICATION :
It is the process of extracting the relevant object/alphabets to from feature vectors. This feature vector is then used by classifiers to recognize the input unit with target output unit [1][3].

(5.) POST PROCESSING : The output of the OCR generally contains error; the debugging responsibility is perform by this phase [1][2].

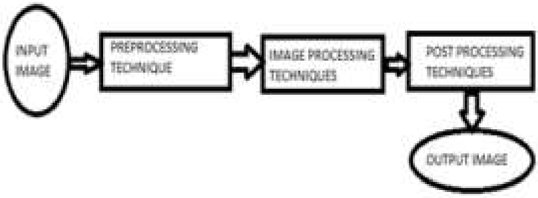

Figure 1:Generalized Procedure of image processing

II.PROCEDURE

OCR is the translator which recognize the text into machine-editable from where the text may be available in the form of handwritten or typed [6]. OCR is used for many applications.

Whenever any text document, which is needed to be edit at that time the document is scanned first then it will be converted into a Bitmap image for removing noise from image, after removing the noise from the image it will be checked for the skew detection where edges will be checked of the document, now the very important part comes into an action that is segmentation where documents will be segme- nted by line, word and by character. After the segmentation process Feature Extraction process comes into the picture, which will extract the character by its feature after this phase classification will be done from the feature database and training database where error correction process will be done on the Bitmap image, now the image will be in the editable form. This is how the process on the text document is done and the The document will be able to edit, so that further changes or updating may be possible on the document. The documents may be of handwritten , scanned or any picture.

Optical Character Recognition can be done via few phase that can be understood with the point no III, IV, V i.e. where all Phase will describe its functionality its features and future scope for the particular phase as described under.

The flow of the process is as shown in the diagram that can be helpful to understand the general procedure of the Optical – Character Recognition system .

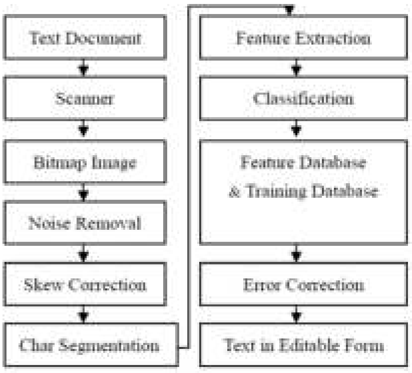

Figure 2: Procedure of General OCR [6].

III.SEGMENTATION

In computer vision, segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics.

The result of image segmentation is a set of segment that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in region is similar with respect to some characteristic or computed property, such as colour, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic. When applied to a stack of images, typical in Medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of Interpo- lation algorithms like marching cubes.

Segmentation can be done by line segmentation, word and by Characters that is as shown in the figures.

In line segmentation as we can see the complete image is coverd by line segmentation process, segmentation is applied and the image is segmented in the line by line that will be further segmented for the word and then for the characters.

Line segmentation is done by scanning the whole image horizontally. Frequency of black pixels in each row is counted in Order to construct the row histogram [4]. The result of the line segmentation is we can see in the figure.

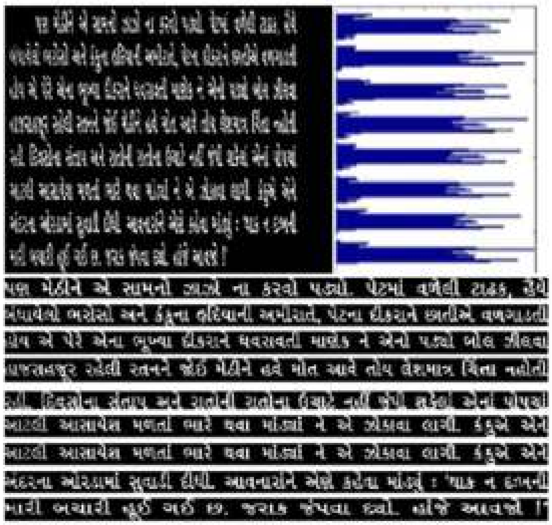

Figure 3: Line Segmentation [4]

Word Segmentation can be handle by the column histogram [4]. The portion of the line with continuous black pixels is consider as the word in that particular line. if no black portion is found then it will be consider as the space between that word. Word segmentation can be

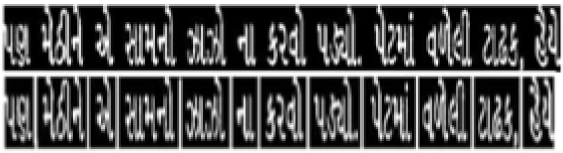

Figure 4: Word Segmentation [4]

Character segmentation is generalized from column histogr- am. Frequency of black pixels in each column is counted in order to construct the column histogram.The position between two consecutive characters, where the number of pixels in a column is zero denotes a boundary between the characters that is as shown in the figure 5.

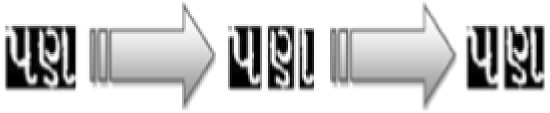

Figure 5: Character Segmentation [4]

Some of the practical applications of image segmentation are:

1. Medical imaging [7]
 1. Locate tumours and other pathologies
 2. Measure tissue volumes
 3. Computer-guided surgery
 4. Diagnosis
 5. Treatment planning
 6. Study of anatomical structure
2. Locate objects in satellite images
3. Face recognition
4. Traffic control systems [8]
5. Brake light detection
6. Machine vision [8]

Several proprietary software packages are available for perfor- ming image segmentation.

1. Pac-n-Zoom Color has a proprietary software that colour segments over 16 million colors at photographic quality.
2. Turtle Seg is a free interactive 3D image segmentation tool. it supports many common medical image file formats and allows the user to export their segmentation as a binary mask or a 3D surface.

Several open source software packages are available for performing image segmentation

3. ITK - Insight Segmentation and Registration Toolkit.
4. ITK-SNAP is a GUI tool that combines manual and semi-automatic segmentation with level sets.
5. OpenCV is a computer vision library originally developed by Intel.
6. GRASS GIS has the program module is map for image segmentation
7. Fiji - Fiji is just ImageJ, an image processing package which includes different segmentation plug-ins.
8. AForge.NET - an open source C# framework.

IV. FEATURE EXTRACTION

Different characters have different features, on the basis of these features the characters are recognized. Thus feature extr- ctioncan be defined as the process of extracting differentiating features from the matrices of digital character. [1]. After char- acter segmentation, each character is processed through a featurizzation routine where the best-describing features will be extracted.

It involves simplifying the amount of resources required to describe a large set of data accurately. When performing ana- lysis of complex data one of the major problems steps from the number of variables involved. Through Feature extraction method 97% accuracy is obtained [2].

It extracts the features of symbols. Features are the charact- eristics. In this, symbols are characterized and unimpor- tant attributes are left out.The feature extraction technique does not match concrete character patterns, but rather makes note of abstract features present in a character such as intersections, open spaces, lines, etc.Feature extraction is concerned with the representation of symbols. The character image is mapped to a higher level by extracting special characteristics of the image in the feature extraction phase [3].

The basic featurization routines were developed and can be used for feature extraction.

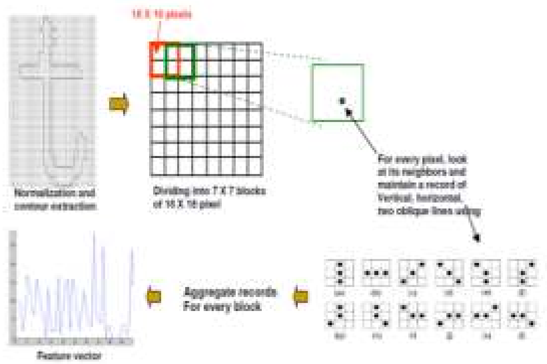
18 X 18 pixels
Normalization and contour extraction
Dividing into 7 X 7 blocks of 18 X 18 pixel
For every pixel, look at its neighbors and maintain a record of Vertical, horizontal, two oblique lines using
Aggregate records For every block
Feature vector

Figure 6: Direction Feature Extraction [9]

Template initialization process is done each of the character. It will be first resize the character in to the 32x32 probabilistic template [9].

Moment descriptors have been studied for image recognition and computer vision since 1960. that overcome the shortcoming of information.and reduces the redundancy [9].

Directional features is for recording the relative neighbouring Pixel positions for each contour pixel and generate a feature vector using that information [9].

V. CLASSIFICATION

During the testing phase, after character segmentation and featurization, each feature vector is classified to one of the trai- ned classes [9]. The results classifications the last stage where we train the neural net using the feature vectors obtained during feature extraction method against the required targets [3]. To optimize the whole recognition process many combina- tional method has been derived.

(1.)Template Matching

Awarding probabilities when template pixel matches with the corresponding pixel in a candidate character image and penali- zing otherwise, forms the core objective of template matching. The template which has the best matc is considered to be the class of the character image. The candidate character image is binary, while the pixel values of the template map g(x, y) are in a range [0,Ninst], therefore g(x,y) is first normalized. The simi- larity of a character image f(x,y) and a template gb(x,y) is defi- ned as a weighted similarity as [9]

$$S_w(f,g) = 1.0 - \frac{1}{N^2}\sum_{x=1}^{N}\sum_{y=1}^{N} w(x,y)|f(x,y) - g_b(x,y)|$$

where the weight w(x,y) is defined as [9]:

$$w(x,y) = \begin{cases} 1.0 & \text{if } g_b(x,y) \text{ is background} \\ g(x,y)/N_{inst} & \text{if } g_b(x,y) \text{ is foreground} \end{cases}$$

(2.)Hierarchy classification :

Kanji and South-East Asian scripts have a large set of alphabets. Hence, one-stage discrimination doesn't generally suffice. In this approach, two – stage classification was used: rough and classification. The aim of rough classification is to cluster copy looking characters into groups and then perform fine classific- ation to extract the right class [9].

VI. CONCLUSION

In this paper we have presented a survey and analysis of various Phase of OCR systems to recognize the general scripts. A lot of work has been done in this field. Still research is going on to improve the efficiency and accuracy for the segmentation, Feature Extraction and classification techniques.

REFERENCE [1] Rohit Verma and Dr.Jahid Ali, "A-Survey of Feature Extraction and Classification techniques in OCR Systems." Proceeding of the international journal of Computer Application and Information Technology, Volume 1, Issue 3, November 2012. | [2] Hemlata Khatri, "Optical Character Recog nition using Segmentation and Feature Extraction." Proceeding of the international journal of Mathe-matics, Science, Technology asnd Management, Volume 2, Issue 2, (ISSN :2319-8125) | [3] Om Prakash Sharma, M.K.Ghose, Krishna Bikram Shah, Benoy Kumar Thakur, "Recent Trends and Tools for Feature Extraction in OCR Technology." Proceeding of the international journal of Soft Computing and Engineering, Volume 2, Issue 6, January 2013. | [4] Prachi Solanki, MalayBhatt, "Printed Gujarati Script OCR using Hopfield Neural Network." Proceeding of the international journal of Computer Applications, Volume 69, Issue 13, May 2013. | [5] Ravina Mithe, Supriya Indalkar, Nilam Divekar, "Optical Character Recognition." Proceeding of the International journal of Recent Technology and Engineering, Volume 2, Issue 1, March 2013 | [6] Monika Pathak, Sukhdev Singh, "Implications and Emerging Trends in Digital Image Processing." Proceeding of the International journal of Computer Science and Information Technologies, Volume 5, Issue 2 2014 | [7] Dzung I. Pham, Chenyang Xu, Jerry L. Prince, "A Survey of Current Methods in Medical Image Segmentation." Proceeding of the journal of Biomedical Engineering in January 1998. | [8] Mr. Mahendra Kumar Pradhan and Mr. Nitin Jain, "Authentication And Verification of Secure Parking using OCR Technique.Proceeding of the journal of the Advance in Electronic and Electric Engineering, Volume 4, Issue 2 , November 2014. | [9] Mudit Agrwal, "Re-Targetable OCR with Intelligent Character Segmentation." In December 2007.