***Recognition of Off-line Hand printed English***

***Characters, Numerals and Special Symbols***

**Abstract— The generic process of Optical Character**

**Recognition (OCR), an area of intensive research in the field of**

**Artificial Intelligence, Pattern Recognition and Computer Vision,**

**aims to recognize text from scanned document images, where**

**data can be in machine printed or hand written format. Optical**

**Character Recognition can improve the interaction between man**

**and machine in various applications including data entry, office**

**automation, digital library, banking applications, health**

**insurance and tax forms etc. Much of work has been done in the**

**recognition of machine printed characters in various languages**

**with considerably good efficiencies, however making robust**

**recognition engines that can be put to recognize hand written and**

**hand printed data with commendable recognition rates still**

**remains as an active area of research owing to the challenges like**

**diverse human handwriting style, variation in shape, angle and**

**style of characters. Taking into account the challenges and scope**

**for improvement in this domain, the work of off-line character**

**recognition of hand printed document images containing English**

**Characters-Uppercase and Lowercase, Numerals and Special**

**Characters has been presented. Statistical, Geometric and**

**Directional Feature Extraction techniques have been applied**

**over segmented character image. Classification was done using**

**Multilayer perception neural network (NN) with back**

**propagation and Support vector machine (SVM) classifier. The**

**recognition rates achieved were up to 98% for Numerals, 96.5%**

**for Special Characters, 95.35% for Uppercase Characters, 92%**

**for Lowercase Characters. The system for combined data setCharacters, Numerals and Special Symbols resulted out to be**

**92.167% accurate, using SVM as classifier.**

***Keywords— — Hand printed character recognition, hand***

***printed numeral recognition, Statistical, geometric and topological***

***features, neural network classification, SVM classifier.***

**I.INTRODUCTION**

**Optical Character Recognition aims at**

**identifying characters in images of printed or handwritten text,**

**in order to encode the text in a format more convenient to**

**edit. Recognition of cursive text has been an active area of**

**research, due to the challenges faced during recognition**

**process as the process incurs high uncertainty in the input**

**documents as writing styles may vary abruptly depending on**

**the interpersonal and intrapersonal variations.. Therefore, it**

**stands out to be a challenging task to devise an OCR system**

**for handwritten document image. Noise, broken characters,**

**touching characters and inappropriate scanning induces further**

**challenges for higher recognition rates. Special Applications**

**and systems include Hand written formula recognition, Bank**

**check analysis and recognition and Information retrieval,**

**Forms Processing, OMR Sheet Processing.**

**The objective is to develop an offline OCR system to**

**recognize Hand printed English Characters, Numerals and**

**Special Characters from the document images, which**

**comprises of text in hand printed format, which would be**

**converted into editable form. Hand printed document image**

**implies that input image comprises of mono-spaced characters**

**whereas in handwriting we have characters that connect,**

**whereas Hand printed input document has more or less**

**uniform height or width.. Hand printed document images also**

**imply that there must be present uniform base-line character**

**images (same horizontal base-line).**

**There can be two modes of recognition, namely off-line**

**and on-line [16] for obtaining input document image. On-line**

**document image recognition implies to storing the character**

**image as a function of time dynamically as the character is**

**being drawn electronically; hence the spatio-temporal**

**information [16] such as order of strokes made by the writer,**

**information about pressure and angle of the pen is readily**

**available. In case of off-line document where acquisition is**

**done prior to recognition, the spatio-luminance [16] of the**

**image is analysed. Therefore, more challenges would be**

**present while recognizing documents in offline mode since we**

**have only static information about the document.**

**An OCR system comprises of different phases as Data**

**Acquisition, Pre-processing, Feature extraction, classification**

**and post-processing [Fig. 1]. Pre-processing includes noise**

**cleaning, skew correction, binarization, segmentation and**

**normalization techniques. In feature extraction phase a set of**

**useful properties of a character are extracted. Statistical,**

**Directional, Topological and Geometric features were**

**extracted to uniquely identify each character. Based on these**

**properties character is assigned to a class in the classification**

**phase. Classification phase involves two major steps -training**

**and testing. Neural Network and Support Vector Machine**

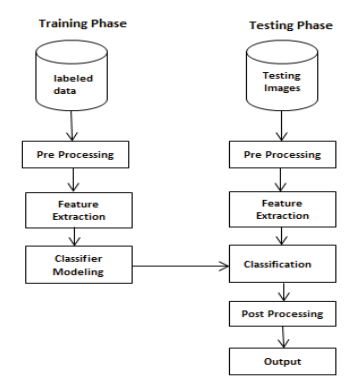
**were used independently for the experiment at the classifier**

**level. Post processing has been applied on the classified**

**characters to generate text output according to input character**

**sequence. Fig. 1 shows the major steps involved in the**

**Approach.**



**Section II of this paper discusses the Data Collection,**

**various Pre-processing steps performed before the data is**

**forwarded to feature extraction. Section III discusses the**

**feature extraction techniques used to extract features which**

**can uniquely identify each character. Section IV discusses the**

**neural network and SVM classification of the characters along**

**with post processing. Section V and Section VI discuss the**

**Testing Results and the conclusion of the approach used**

**respectively.**

**II. DATA COLLECTION AND PRE-PROCESSING**

**A. Data Set Preparation**

**The data set was obtained by taking written samples of 40**

**people, where each writer had written a document containing**

**instances of each of the 26 characters, 10 numeral digits or 8**

**special characters in the range-[10-30].These documents were**

**scanned at resolution-200 and 300 dpi. There were nearly**

**3517, 2340 and 1804 instances of Upper case characters,**

**Lowercase characters and Numerals respectively. The**

**collected data has been divided into the ratio of 60:40 for the**

**training and testing purpose. However, these partitions were**

**changed iteratively to test the system for best accuracy and for**

**obtaining the optimal data set. The digitized images were**

**saved in BMP format. Fig. 2, Fig. 3 and Fig. 4 show samples**

**of acquired images for hand printed characters (Uppercase and**

**Lowercase) and Numerals, respectively.**

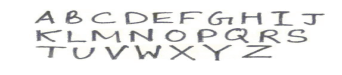
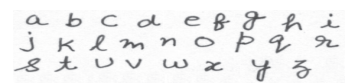


Fig. 2. Sample Images of Hand printed Scanned characters



**Fig. 3. Sample Images of Hand printed Scanned Lowercase Characters**

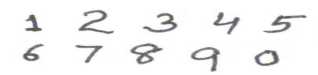


Fig. 4. Sample Images of Hand Printed Scanned Numerals



Fig. 5. Sample Images of Hand printed Scanned Special Symbols

**B. Pre-processing**

**Pre-Processing enhances a document image preparing it**

**for the feature extraction phase in the OCR system. In order to**

**achieve high recognition rate, prior to character recognition, it**

**is essential to eliminate the noise and imperfections introduced**

**in the image. The three techniques adopted for pre processing**

**in this approach were Binarization, Thinning and Size**

**Normalization. In Binarization, grey-scale image is converted**

**into binary image with the help of thresholding. The Otsu**

**method was applied to convert the image into binary image.**

**Otsu’s method selects the threshold by minimizing the withinclass variance of the two groups of pixels separated by the**

**thresholding operator.**

**The technique of Thinning converts the character images**

**to single pixel width. The binary image is represented as ones**

**and zeros. “1” is used to represent object pixel and “0” is used**

**to represent background. Zhang-Suen Algorithm [5] was used**

**for this purpose. Segmentation is a process that determines the**

**constituents of an image. It is necessary to locate the regions**

**of the document where data have been printed and distinguish**

**them from figures and graphics. When applied to text,**

**segmentation is the isolation of characters or words. In this**

**approach, Contour tracing was used to segment each character**

**image. Features extracted from the processed image should**

**not be affected by the size of the image. Each Segmented**

**Image was normalized to 20 x 20 pixels dimensions.**

**III. FEATURE EXTRACTION**

**The objective of feature extraction phase is to extract the**

**essential and differentiable characteristics of the symbols.**

**Feature space is much less than input image space as we**

**extract only essential properties for higher recognition rate.**

**Features are classified into the following categories on the**

**basis of methods of extraction.**