CHAPTER 2
LITERATURE REVIEW

The recent studies in the field of computer vision have shown a great amount of interest in text extraction from images. Various techniques have been proposed for detection and recognition of text in the images. These techniques are based on different properties related to text and background of the image, such as, color or intensity of text region with reference to background, connected-components property of text region, presence or absence of edges, etc. These properties are exploited to distinguish text regions from the image background.

The text extraction system works in sequence of steps which involves detection, localization, extraction, and recognition of the text from the image. The text extraction and recognition is a complicated task due to variable text size, font, style, orientation, alignment and complex background. The various algorithms have been reported for the extraction of text from camera captured images, video frame images, printed scanned document, degraded document images and handwritten scanned documents.

In the present research work, thorough literature review has been carried out which is presented into two parts. In first part, the literature deals with text extraction and recognition system for Indian scripts. Second part of literature review covers existing techniques and methods used in development of text extraction system. The critical analysis of literature review in these two parts is included in section 2.1.section 2.2 respectively.

2.1 RESEARCH ANALYSIS ON TEXT EXTRACTION SYSTEM FOR INDIAN SCRIPTS

The text extraction and recognition system for Indian scripts is in its initial phase. The work on few Indian scripts has been reported in literature. Most of the work is found on Machine Printed document Images and handwritten document images of Indian scripts (Devanagari, Gurmukhi, Bangla, Tamil, Telugu, Oriya, Gujarati and Kannada). Few attempts have been noticed on text extraction from Natural scene images in Devanagari and Tamil. In case of Gurmukhi script, Most of

the work has been reported on handwritten online and offline scripts. Very Few attempts have been made on text extraction from natural scene images.

2.1.1 Devanagari Script

Agnihotri et al. [21] proposed Devanagari hand written text recognition system. This system extracts Zone directional feature from character images and convert into chromosome bit string of length 378. The character recognition is carried out by a special function called fitness function which computes the Chromosome difference between unknown character and Chromosome database. The recognition of the system was tested on training set containing 904 characters. The results showed precisions 85.78 %(match) and 13.35 %(mismatch).

Murthy et al. [22] developed text extraction system using 2-D Haar Discrete Wavelet Transformation and K-Means Clustering. Morphological operators were used to distinguish the text and non text regions. The system was tested on 150 images which contained text in Hindi and English.

RajibGhosh et al. [23] introduced feature extraction system for an online handwritten character recognition system for two scripts namely Bengali and Devanagari script. The features of handwritten characters were extracted which were based on unique basic strokes. The stroke information such as, writing direction, slope, curvature, curliness, standard deviation of x and y coordinates features were considered for feature extraction. The recognition accuracy of the proposed system is 78% for Bengali script and 77.23% for Devanagari script.

Hasan, et al. [24] proposed a system to automatically detection and recognition of Bengali text on road signs and translate into voice stream. The system contained three modules: text detection and extraction, optical character recognition and speech synthesis. First module included data (i) acquisition, (ii) pre-prcrossing (iii) text detection and localization. The OCR module includes (i) character segmentation (ii) features extraction and (iii) character classification using artificial neural networks. The speech synthesis module converts the recognized text into voice streams. The system in this study has been tested with several chunks of printed Bangla texts or textual images (.bmp/.jpg file) and observed the system outputs. The experimental results reveal that 820 total numbers of characters of Sulekha font with size gave 96% accuracy. Similarly 753 total numbers of characters

of SutonnyMJ font with size 24 gave 98% accuracy and 669 total numbers of characters of Moina font with size 20 gave 97% accuracy.

Sethi et al. [25] used structural approach to recognize number of Devanagari script. The segmentation techniques were based on horizontal and vertical segmentation where skew correction was carried for right and left slants. The decision tree was used as analysis tool to access the presence/absence of these primitives and their interconnection

Veena Bansal et al. [26] proposed OCR for Devanagari script. Various character features such as Vertical bar feature, Horizontal zero crossings, Number of positions of the vertex points were considered for recognition. The character classification was carried by tree classifier. The testing results showed 93% accuracy.

2.1.2 Gurmukhi script

G S Lehal et al. [27] developed a system for recognition of machine printed Gurmukhi script. The system performed the segmentation in three successive stages namely line segmentation, word segmentation and character segmentation. The horizontal and vertical projection profiles were used for segmentation of line and words respectively. The different structural features of Gurmukhi script such as Number of junctions with the headline, Presence of Sidebar, Presence of a loop, Aspect Ratio were considered for recognition. The overall recognition rate of the proposed system reported as 96.6 % with speed of 175 character/ seconds. M. K. Jindal et al. [28] proposed new algorithm to segment touching Gurmukhi characters in middle zone. They prepared a database of poor quality scanned printed documents which contained around 30% touching characters. The different categories of touching characters were identified which were based on structural features of Gurmukhi characters. About 500 scanned documents which contained 6000 touching characters were considered for testing. The proposed algorithm successfully segment characters which contained 92-95% touching characters and showed recognition rate of 91%.

U. Bhattacharya, et. al [29] have developed a system to extract Devanagari and Bangla text from natural scene images. One assumption was made to keep character size sufficiently large and or thick. A global binarization method like the

well-known Otsu's technique has been used. Otsu binarization method was used twice to get the desired result. The connected components approach was used to capture headlines. The morphological opening operations Erosion followed by dilation on each connected component to extract possible horizontal line segments was used. The precision and recall values of the algorithm obtained on the basis of the set of 100 images were respectively 68.8% and 71.2%. The proposed algorithm worked well even on slanted or curved text components of Devanagari and Bangla. It failed whenever the size of such curved or slanted text was not sufficiently large.

Sharma et al. [30] proposed Gurmukhi recognized system for isolated handwritten Gurmukhi using Neocognitron. The Neocognitron defined a hierarchical multilayered neural network used for classification. Total 15000 character images were considered for testing. The learned images have showed 91.77% accuracy whereas unlearned images showed 93.79% accuracy. Mangla et al. [31] proposed segmentation algorithm for touching and broken characters in handwritten Gurmukhi. The various challenges faced during segmentation were well described and analyzed. The proposed method for segmentation is based on neighboring pixels analysis. The result revels 95% accuracy for Gurmukhi touching and broken characters.

Budhiraja et al. [32] proposed an OCR system for isolated Gurumukhi characters. The character features were extracted from test sample images of isolated handwritten character. Various features used for recognition are Zoning, Projection histogram, Distance profile features, and Background directional features. Support Vector Machines classifier was used and obtained 95.04% of recognition accuracy.

2.1.3 Kannada

T. K. Boaz et al. [33] introduced method for localisation of Kannada text from video frames. The method was based on multiple frames interrogation approach in which frames were integrated on the basis of characteristic like character location, edge distribution and pixel contract. The Rober's edge detection was used to create edge image and morphological operators were used to filter text and non text regions. The experimental result showed 92.00 % accuracy. Vipin Gupta et al. [34] proposed frame work to extract Kannada text from camera captured image. The method was also based on edge detection, color clustering and

morphological operators. Ramappa et al. [35] have proposed features and classification method to recognise handwritten Kannada numerals. They have used eight different features namely, run length count and carvel, directional chain code, image fusion, radon transform, fan beam projections, discrete Fourier transform and zone extraction. The image dataset considered consist of 1000 test images. The given set of image was converted into gray scale and later into binary images using global thresholding methods. The thinning technique was also applied on binary images to extract the different features. The various classifiers like Euclidean distance, Cosine distance, K-NN, K-means, Classifier fusion, etc were used for recognition. Among the various features the directional chain code feature gave better result. The overall recognition rate with using multiple classifiers was 98.6%.

2.1.4 Oriya Script

Chaudhuri et al. [36] proposed an Optical Character Recognition system for printed Oriya script. In pre-processing phase, skew detection and line segmentation were used. To extract character features combined features were used which includes stroke and run-number based features and features from water reservoir. The overall word segmentation accuracy of the system was 97.7%. Roy et al. [37] presented off-line Oriya handwritten numerals recognition system based on neural network based classifier. Overall accuracy of the system was 94.81%.

2.1.5 Malayalam Script

Rajashekararadhya et al. [38] proposed feature extraction method for offline Malayalam hand written number system. Zoning based hybrid feature was extracted from character images. The zoning feature was extracted by computing the centriod of the character image and divides the image into fifty equal parts. The zone features were computed by measuring average distance of each pixel with reference to the centroid. The empty zones assigned zero in feature vector. The zoning features are used for classification and recognition using K-NN (K-nearest neighbor classifier) and feed forward back propagation neural network. The proposed system in the study was tested on 2000 Kannada numeral samples. Experiment results raveled 99% recognition.

2.1.6 Telugu Script

Rajasekaran et al. [39] proposed printed Telugu character recognition method which worked in two stages. In the first stage, a syntax-aided recognition scheme was used which removed primitive shapes from Telugu characters. In the second phase, Decision tree based classification approach was used to recognise basic shape. Sukhaswamy et al. [40] proposed text extraction system for Telugu script using Multiple Neural Network Associative Memory. In this system, the exemplars were divided into groups and each such group was trained into a separate network. The character pattern to be recognized was placed to each of these networks. The limitations this approach was that the character recognition was not invariant of character size, translation and rotation. Negi et al. [41] introduced an OCR for Telegu script using template matching, fringe distance and testing results showed 92% accuracy.

Most of work reported in Indian Language were on hand written and document OCRs. Table 2.1 shows brief findings of current research work on Indian script.

Table2.1: Highlight of the work done on Indian script text extraction system

Author	Language	Approach Feature	Result analysis
G S Lehal et al.	Gurmukhi	Horizontal and Vertical Projection Profiles, KNN and SVM classifier	reported as 96.6 % with speed of 175 character/ seconds
M. K. Jindal et al.	Gurmukhi	algorithm to segment touching Gurmukhi characters, structured features of Gurmukhi characters	Recognition rate of 91%.
Saini et al.	Gurmukhi and Shahmukhi	Shahmukhi to Gurmukhi transliteration Statistical analysis-character and words n-gram analysis	Average transliteration accuracy of 91.37%.
D.Sharma et al.	Gurmukhi	Neocognitron-hierarchical multilayered neural network based classifier	Recognition accuracy 91.77% (learned image) and 93.79%(un learned image)
Mangla et al.	Gurmukhi	Algorithm for segmentation of touching and broken characters, vertical and horizontal	95% accuracy for Gurmukhi touching and broken characters.

Author	Language	Approach Feature	Result analysis
		projection profiles.	
Singh and Budhiraja	Gurmukhi	Zoning, Projection histogram, Distance profile features, and Background directional features. Support Vector Machines classifier	95.04% of recognition accuracy for isolated characters
T K Boaz et al	Kannada	multiple frames interrogation approach, Rober's edge detection, morphological operators	92.00 % accuracy.
Vipin Gupta et al	Kannada	edge detection, color clustering	Visual observation show satisfactory results
Ramappa et al.	Kannada	eight different features, global thresholding, K-NN, K-means	98.6%. accuracy
Chaudhuri et al.	Oriya	skew detection and line segmentation	97.7% accuracy
Rajashekararadhya et al.	Malayalam	Zoning based hybrid feature, K-NN (K-nearest neighbor), feed forward back propagation neural network	99% recognition
Rajasekaran et al.	Telugu	syntax-aided recognition scheme, Decision tree based classification	96% recognition
Sukhaswamy et al.	Telugu	Multiple Neural Network Associative Memory	Character recognition is not invariant of size, translation and rotation.
Chakravarthy	Telegu	template matching, fringe distance	92% accuracy.

2.2 VARIOUS TECHNIQUES AND METHODS FOR TEXT EXTRACTION AND RECOGNITION

Nobuo Ezaki, et al. [42] developed a system for visually impaired persons. The system was developed to provide assistance to a visually impaired person for walking on the streets. The system enclosed head mounded camera along with text extraction system which provided text information of surrounding area. For example, a "stop" sign at a crossing provide inform that the person is near to crossing area. Similarly the text written on walls, banners may provide information about the surroundings. The Sobel edge detection, Otsu binarization, connected-component extraction based methods were implemented and evaluated. The

effectiveness of the methods was evaluated on data set of the ICDAR 2003 Robust Reading Competition. A group of 55 images that contained only small characters were selected for evaluation. The efficiency of the morphological method on these images was reported as precision 38%, recall 55% and f 47%.The edge based method on these images was also tested. The performance matrix reported is: precision =26%, recall= 48% and f=37%. The results revealed that the morphological method was effective for small characters. Angadi et.al [43] proposed a methodology to detect and extract text regions from low resolution natural scene images. Their proposed work used Discrete Cosine Transform (DCT) based high pass filter to remove and suppress the constant background. The texture feature matrix was computed on every 50x50 block of the processed image. A newly defined discriminate function was used to classify text blocks. The detected text blocks were merged to obtain new text regions. Finally, the refinement phase was a post processing step used to improve the detection accuracy. This phase used to cover small portions of missed text present in adjacent undetected blocks and unprocessed regions. The proposed methodology had been conducted on 100 indoorand outdoor low resolution natural scene images of each of size 240x320 containing text of different size, font, and alignment with complex backgrounds containing Kannada text and English text.

Wumo Pan et al. [44] proposed method for text detection based sparser presentation. Canny edge detection method was used to generate edge image and connected component was used to filter non text regions. Text detection method was tested on 120 natural scene images and 251 images were taken from ICDAR-2003 competition dataset. The experimental results showed precision 67% and recall 75.23%. Yamaguchi et al. [45] introduced digits recognition system to read telephone numbers written on signboards. The candidate text regions in sign boards were extracted using edge detection and connected component methods. Before segmentation, skew detection and correction by Hough transform was also carried out. Matsuo et al. [46] suggested a method to extracts text from scene images using adaptive thresholding. Yang et al. [47] proposed a framework to detection sign from signs board. This framework designed to handle challenges in sign detection, extraction and recognition. The framework proposed a robustly automatic system to

understand sign in natural scene images which can handle different lighting and view angle problems.

Gatos et al. [48] proposed a method based on Binarisation and connected component analysis for detection of text in natural scene images. The binarisation method was able to handle complicated issues of object shadows, non uniform illumination and low color contrast. The text regions were clubbed by using connected component analysis approach. The ICDAR2003 Robust Reading Competition database was used for testing of the binarisation algorithm.

Pan et al. [49] proposed a hybrid method based on confidence map and Conditional Random Field (CRF). The text confidence map was used along binarisation method to generate candidate text region image. The CRF model along with energy minimization models were used to label the text components and filtered non-text regions. Data structure minimum spanning tree was used to express neighbourhood components in edge image to build text region. Shyama et al. [50] projected a color based text segmentation technique to extract text from camera captured image or video frames. Edge detection method and edge enhancement methods were used to find edge image and filter blurred edges. B. Gatos et al. [51] introduces the idea of adaptive zoning features for word or character recognition. The zoning features were extracted by adjusting the position of the zone. The adjustment of position of zone was based on local pattern. The local patterns followed were based on rule of maximizing the local pixel density around the zone. The experimental results showed significant enhancement in character recognition accuracy from 85.98% to 88.35%.

Vijay Kumar et al. [52] proposed a method to extract ports information displayed in video frame. The text regions were cropped from video frame using color histogram and later on Canny edge detection method was applied to generate edge image. The text region fed to Optical Character Recognition (OCR) to produce index able keywords. Xiaopei Liu et al. [53] developed a method to detect and localize text region in the scene images. The method worked in three phases. In first phase, edge detection was implemented by computing magnitude of second derivates of intensity values around edges. In the second phase, clustering technique

was used to localise text region. In the last phase, the detected text regions were tested by professional OCR engine. Seeri et al. [54] proposed a hybrid approach, which locates text in natural scene images with different backgrounds. They have used edge feature using wavelet domain and fuzzy classification to recognize the text. Leon et al. [55] presented technique for caption text detection. The technique incorporated Haar wavelet decomposition to detect candidate text regions and geometric features of caption text for recognition. NitiSyal et al. [56] proposed a hybrid approach to extract from the text in image. The hybrid approach used based N fusion of Daubechies DWT MATLAB function, radiant difference and SVM.

Tokas et al. [57] presented different feature extraction methods to classify the 26 handwritten capital alphabets written by 25 different writers. Analysis of these features extraction methods with back propagation neural network classifier has been done. Neural network was used for classification with most of the feature vector types. It was observed that combining different feature vector into a single feature vector for hybrid method results in higher recognition rate compared to its individual feature extraction method in case of accuracy. Hossain et al. [58] proposed feature extraction method based on projection histogram for Bangla handwritten numerals. Various features extraction methods such as crossings, Fourier transform, and zoning were compared. The experiment was carried out on bangle numeral and achieved 94.12 % accuracy.

Gatos et al. [59] Proposed a technique to detect text from indoor and outdoor images. The technique was used to isolate background using binarisation methods. The binarisation method segment the text region from the background then commercial OCR was used for testing purpose. Tran et al. [60] proposed an approach for finding text in images by using ridges at several scales. The proposed method didn't depend on a particular alphabet. The experimental results showed a good detection.

Jindal et al. [61] proposed method to segment the Gurmukhi touching characters from the middle. For segmentation, various classifications were carried out which were based on structural properties of the characters. The algorithm was tested on degraded Gurmukhi script characters and had shown rescannable enhancement in

segmentation process. The algorithm was capable to segment more than two touching characters in the single Gurmukhi word.

Saini et al. [62] described a corpus based transliteration system for Punjabi language. They developed a new system for the first time of its kind for Shahmukhi script of Punjabi language. The proposed system for Shahmukhi to Gurmukhi transliteration was implemented with various research techniques based on language corpus. The corpus analysis program was run on both Shahmukhi and Gurmukhi corpora for generating statistical data for different types such as character, word and n-gram frequencies. This statistical analysis was used in different phases of transliteration.

Rodrigues et al. [63] proposed the technique to segment the postal code image. The image was segmented by using histogram projection profiles. The image was stored in 2-dimension matrix, and then segmentation algorithm was applied. The decision tree classifier was used to recognition of the segmented characters.

Elena Kozerenko et al. [64] focused on the role of features for the implementation of the transfer-based machine translation systems. The method was based on semantic content of syntactic structures and used for the study of the English and Russian language. The texture analysis was also carried out.Wang et al. [66] studied the light verb constructions in Japanese translation system. The study was conducted to explore the various issues to convert Japanese LVCs into Chinese based on phonemes. The various rules were constructed by observing the experiments on 200 sentences. The pattern-based translation engine was also explored for translation. The overall accuracy of the proposed system in the study was 80%.

Gao et al. [65] presented a new adaptive algorithm for automatic detection of text from a natural scene. The whole algorithm was applied in a prototype system that could automatically detect and recognize sign input from a video camera, and translate the signs into English text or voice streams. The optimization algorithm was applied enhancing the performance of the color modeling technique and various constrains of text layout were handled. The algorithm proposed in the study was a hybrid one which was a combination of various approaches used for text detection

and localization. Table2.2 shows highlight of the work done on various methods and techniques used in text extraction system.

Table 2.2: various methods and techniques used in text extraction system

Author	Language	Approach/ Features	Result analysis
Nobuo Ezaki	English	Sobel edge detection, connected components	Precision 26%, recall 48% and f 37%.
Angadi et.al	Kannada text, English text	Discrete Cosine Transform,	Accuracy rate of 96.6%
Wumo Pan et al.	English	sparser presentation, Canny edge,	Precision 67% and recall 75.23 %.
Yamaguchi et al	English	edge detection and connected component methods	Digit extraction rate of 99.2% and a correct digit recognition rate of 98.8%.
Gatos et al.	English	Binarisation, connected component analysis,	more than 50% improvement of the FineReader_6
Pan et al.	English	Conditional Random Field (CRF). Edge detection	Precision67.64% and recall75.23%
VijayaKumar et al.	English	Canny edge detection, color histogram	Accuracy 79%
Seeri et al.	English	High pass filter	P= 79.54% ,recall 89.21%
Angadi et al	English	DCT based high pass	Accuracy 96.6%
Anuj et al.	English	Connected component based and edge based algorithm	Edge-based-methods: Precision 47.4% and recall rate is 75.09%, connected component algorithm, precision 50.10% and recall rate is 73.42%