Improve Optical Character Recognition Using Templates & Correlation

Sukhpreet Singh
YCOE Talwandi sabo

Abstract— **This paper presents a new improved technique on English OCR techniques. English OCR system is compulsory to convert numerous published books of English into editable computer text files. Latest research in this area has been able to grown some new methodologies to overcome the complexity of English writing style. Still these algorithms have not been tested for complete characters of English Alphabet. Hence, a system is required which can handle English text and identify characters. The propose algorithm is verified by considering various noises like salt and pepper, Gaussian and Poison and it found that proposed algorithm along with bilateral filter provide better results The proposed algorithm used templates and use correlation to match the letters. Developed algorithm is designed and implemented in MATLAB and tested on various images. The proposed strategy has achieved 97.3 % (average) accuracy which was 94.33 % in Huang and Lin (2012) [16].**

***Index Terms*: OCR, Bilateral Filter, Correlation.**

I. INTRODUCTION

Optical Character Recognition [1] – [5] is a process that can convert text, present in digital image, to editable text. It allows a machine to recognize characters through optical mechanisms. The output of the OCR should ideally be same as input in formatting. The process involves some pre-processing of the image file and then acquisition of important knowledge about written text.

That knowledge or data can be used to recognize characters. OCR [1] is becoming an important part of modern research based computer applications. Especially with the advent of Unicode and support of complex scripts on personal computers, the importance of this application has increased.

The current study is focused on exploration of possible techniques to develop an OCR [2] system for English language when noise is present in the signal. A detailed analysis of English writing system has been done in order to understand the core challenges. Existing OCR systems are also studied to know the latest research going on in this field. The emphasis was on finding workable segmentation technique and diacritic handling for English strings, and built a recognition module for these ligatures. The complete methodology is proposed to develop an OCR system for English and a testing application is also made. Test results are reported and compared with the previous work done in this area.

II. DESIGN OF OCR

Various approaches used for the design of OCR systems are discussed below:

Matrix Matching [6]: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

Fuzzy Logic [6]: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/ white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

Feature Extraction [3]-[6]: This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

Structural Analysis [6]: Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

Neural Networks [6]: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

III. RELATED WORK

Claudiu et al. (2011) [1] has investigated using simple training data pre-processing gave us experts with errors less correlated than those of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates.

ISSN: 2278 – 1323

Georgios et al. (2010) [2] has presented a methodology for off-line handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at each iteration have balanced (approximately equal) numbers of foreground pixels, as far as this is possible.

Sankaran et al. (2012) [3] has presented present a novel recognition approach that results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably good performance on good quality documents, but fail easily in presence of degradations. Also, classical OCR approaches perform poorly over complex scripts such as those for Indian languages.

Jawahar et al. (2012) [4] has propose a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional Long- Short Term Memory (BLSTM).

Zhang et al. (2012) [5] has discussed the misty, foggy, or hazy weather conditions lead to image color distortion and reduce the resolution and the contrast of the observed object in outdoor scene acquisition. In order to detect and remove haze, this article proposes a novel effective algorithm for visibility enhancement from a single gray or color image.

Sankaran et al. (2012) [16] has proposed a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional LongShort Term Memory (BLSTM). Our approach does not require word to character segmentation, which is one of the most common reason for high word error rate.

Hem et al. (2012) [7] has discussed that Optical Character Recognition (OCR) is a type of document image analysis where scanned digital image that contains either machine printed or handwritten script input into an OCR software engine and translating it into an editable machine readable digital text format. A Neural network is designed to model the way in which the brain performs a particular task or function of interest. Each image character is comprised of 30×20 pixels.

Yang et al. (2012) [8] has proposed a novel adaptive binarization method based on wavelet filter is proposed in this paper, which shows comparable performance to other similar methods and processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. The proposed method is evaluated on complex scene images of ICDAR 2005 Robust Reading Competition, and experimental results provide a support for our work.

Sumetphong et al. (2012) [9] has proposed a novel technique for recognizing broken Thai characters found in degraded Thai text documents by modeling it as a set-partitioning problem (SPP). The technique searches for the optimal set-partition of the connected components by which each subset yields a reconstructed Thai character.

AlSalman et al. (2012) [10] has proposed that Braille recognition is the ability to detect and recognize Braille characters embossed on Braille document. The result is used in several applications such as embossing, printing, translating...etc. However, the performance of these applications is affected by poor quality imaging due to several factors such as scanner quality, scan resolution, lighting, and type of embossed documents.

Mutholib et al. (2012) [11] has proposed that Android platform has gained popularity in recent years in terms of market share and number of available applications. Android operating system is built on a modified Linux kernel with built-in services such as email, web browser, and map applications. In this paper, automatic number plate recognition (ANPR) was designed and implemented on Android mobile phone platform.

Chi et al. (2012) [12] has proposed that because of the existence of possible carbon and seals, it's quite often that images of financial documents such as Chinese bank checks are suffered from bleed-through effects, which will affect the performance of automatic financial document processing such as seal verification and OCR.

Ramakrishnan et al. (2012) [13] has proposed that many machine learning algorithms rely on feature descriptors to access information about image appearance. Using an appropriate descriptor is therefore crucial for the algorithm to succeed. Although domain- and task-specific feature descriptors may result in excellent performance, they currently have to be hand-crafted, a difficult and time-consuming process.

Chattopadhyay et al. (2012) [14] has worked on a low complexity video OCR system has been presented, that can be deployed on an embedded platform. The novelty of the proposed method is the use of low processing cycle and memory and yet getting a recognition accuracy of 84.23% which is higher than the usual video OCR recognition accuracy.

Malakar et al. (2012) [15] has described that extraction of text lines from document images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In case of handwritten document images, presence of skewed, touching or overlapping text line(s) makes this process a real challenge to the researcher.

Huang and Lin (2012) [16] has studied that the major drawback of the existing methods is the long training time. Speeding up the training time for most techniques

ISSN: 2278 – 1323

usually suffers the high reduction on the recognition rate. In this work, we present a neural network based approach to largely reduce the training time while maintain the high recognition rate.

IV. PROPOSED ALGORITHM

This section will explain the proposed algorithm, i.e. what are different steps involve for achieving the OCR recognition. Figure 1 is showing the different steps required to do correlation based OCR recognition

Step 1: As a first step, the image is cropped to fit the text.
Step 2: Once trimmed the image, the next step is to separate each line.
Step 3: Once attained unconnectedly each line of the image, we continue to extract one letter of the image matrix and continue until all word get separated.
Step 4: *Classification*
The main process is used for the classification was the two-dimensional correlation. This process gives a value of the correspondence between two matrices (images).
Step 5: Output File

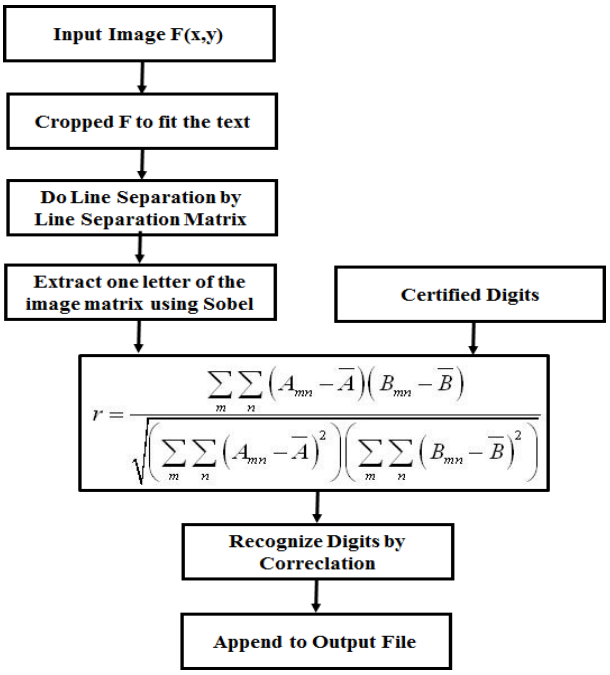

Fig. 1 Proposed Algorithm Flowchart

5. Experimental setup & Experimental results

Table 1: Input data set

Serial no	Name	Format	Total characters
1	a1	.jpg	48
2	a2	.jpg	16
3	a3	.jpg	76
4	a4	.jpg	52
5	a5	.jpg	112
6	a6	.jpg	50
7	a7	.jpg	80
8	a8	.jpg	83
9	a9	.jpg	84
10	a10	.jpg	100
11	a11	.jpg	109
12	a12	.jpg	71
13	a13	.jpg	71
14	a14	.jpg	87
15	a15	.jpg	68
16	a16	.jpg	31
17	a17	.jpg	79
18	a18	.jpg	86
19	a19	.jpg	56
20	a20	.jpg	111

Table 1 is showing the input table for the experimental purpose. Different number plate images are selected to verify the proposed algorithm. Propose algorithm is designed and implemented in MATLAB and the Table 1 is passed to it for verification.

By implementing the proposed algorithm and Huang and Lin (2012) [16] algorithm on the data set shown in table 1 we have taken the hit and miss rate from which we have evaluated the accuracy and error rate. Accuracy and error rate shows that the proposed algorithm provide better results. Following figures demonstrate the performance of the proposed algorithm.

Performance analysis

Table 2 shows the analysis hit and miss analysis of the proposed algorithm by considering the bilateral filter. Hits are incremented when we found accurate digit otherwise miss will be incremented.

Table 2: Hit and Miss Analysis of proposed algorithm with bilateral filter

Name	Total characters	Hits	Miss
a1	48	48	0
a2	16	16	0
a3	76	74	2
a4	52	47	5
a5	112	112	0
a6	50	49	1
a7	80	72	8
a8	83	80	3
a9	84	84	0
a10	100	100	0
a11	109	99	10
a12	71	70	1
a13	71	71	0
a14	87	85	2
a15	68	68	0
a16	31	30	2

ISSN: 2278 – 1323

a17	79	77	2
a18	86	82	4
a19	56	54	2
a20	111	106	5

Table 3 shows the analysis hit and miss analysis of the proposed algorithm by not considering the bilateral filter. It is shown in table 3 that the hits are decremented in some of the data samples.

Table 3: Hit and Miss Analysis of proposed algorithm without bilateral filter

Name	Total characters	Hits	Miss
a1	48	46	2
a2	16	16	0
a3	76	74	2
a4	52	45	7
a5	112	103	9
a6	50	48	2
a7	80	72	8
a8	83	75	8
a9	84	82	2
a10	100	100	0
a11	109	97	12
a12	71	70	1
a13	71	69	2
a14	87	84	3
a15	68	58	10
a16	31	26	5
a17	79	75	4
a18	86	82	4
a19	56	54	2
a20	111	99	12

Table 4 is showing the accuracy analysis of the proposed algorithm without bilateral filter as well as with bilateral filter. It is evidently shown that the proposed algorithm with bilateral filter provide better results than without bilateral filter.

Table 4: Accuracy analysis of Without Bilateral Filter and with Bilateral Filter

Name	With Bilateral Filter	Without Bilateral Filter
a1	100	95.8
a2	100	100
a3	97.3	97.3
a4	90	86
a5	100	91.9
a6	98	96
a7	90	90
a8	96.4	90
a9	100	97
a10	90.9	90.9
a11	98.6	89
a12	100	98
a13	100	97
a14	97.7	96
a15	100	85
a16	96.8	83
a17	97.5	95
a18	95.4	95.4
a19	96.43	96.43
a20	97.9	89

Table 5 is showing the error rate analysis of the proposed algorithm without bilateral filter as well as with bilateral filter. It is evidently shown that the proposed algorithm with bilateral filter provide lower error rate than without bilateral filter.

Table 5: Error rate analysis of Without Bilateral Filter and with Bilateral Filter

Name	With Bilateral Filter	Without Bilateral Filter
a1	0	4.2
a2	0	0
a3	2.7	2.7
a4	10	14
a5	0	9.1
a6	2	4
a7	10	10
a8	3.6	10
a9	0	3
a10	0	0
a11	1.4	11
a12	0	2
a13	0	3
a14	2.3	4
a15	0	15
a16	3.2	17
a17	2.5	5
a18	4.6	4.6
a19	3.57	3.57
a20	2.1	11

Table 6 is showing the computation time analysis of the proposed algorithm without bilateral filter as well as with bilateral filter. It is evidently shown that the proposed algorithm with bilateral filter take much more time than the without bilateral filter. So one can say that the proposed algorithm with bilateral filter is time consuming than without bilateral filter.

Table 6: Computational time analysis of Without Bilateral Filter and with Bilateral Filter

Name	With Bilateral Filter (sec's)	Without Bilateral Filter(sec's)
a1	18.03	5.82
a2	16.53	3.49
a3	21.55	3.10

ISSN: 2278 – 1323

a4	14.79	3.91
a5	32.83	8.63
a6	14.19	2.08
a7	39.15	6.17
a8	33.59	5.39
a9	28.06	4.29
a10	46.72	5.50
a11	52.41	8.64
a12	20.33	3.15
a13	37.27	4.09
a14	23.93	5.24
a15	24.25	5.04
a16	20.37	1.66
a17	22.67	4.09
a18	24.85	6.22
a19	22.46	2.88
a20	41.35	6.57

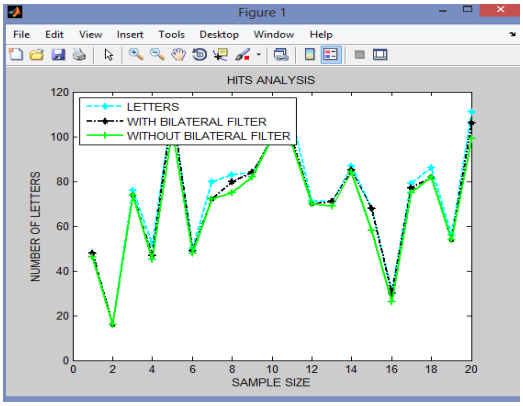

Figure 6.1 Hits analysis

Figure 6.1is showing the hits analysis of the proposed algorithm by considering the bilateral filter as well as without using bilateral filter. Figure 6.1 clearly demonstrate that the hits are either equal in the both techniques and also some time hits are quite better when we are using the bilateral filter. In figure 6.1 the black line shows the proposed algorithm with bilateral filter and also green line show the proposed algorithm without using the bilateral filter.

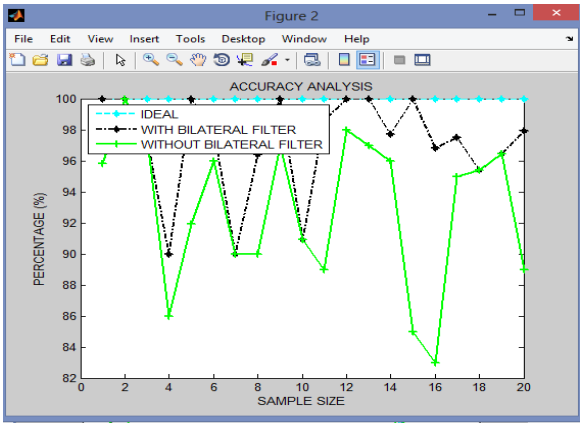

Figure 6.2 accuracy analysis

Figure 6.2is screening the accuracy analysis of the proposed algorithm by considering the bilateral filter as well as without using bilateral filter. Figure 6.2 clearly express that the accuracy is either equal in the both techniques and also some time accuracy are quite better when we are using the bilateral filter. In figure 6.2 the black line shows the proposed algorithm with bilateral filter and also green line show the proposed algorithm without using the bilateral filter.

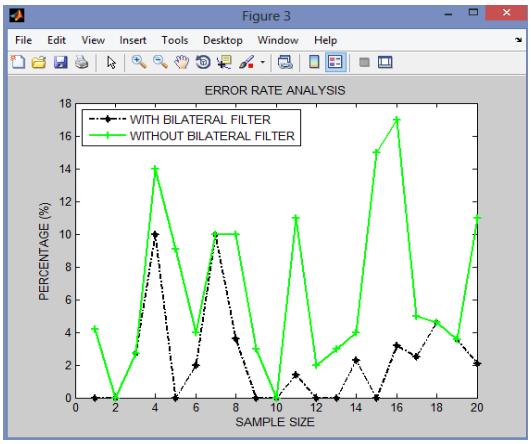

Figure 6.3 Error rate analyses

Figure 6.3is viewing the error rate analysis of the proposed algorithm by considering the bilateral filter as well as without using bilateral filter. Figure 6.3 clearly express that the error rate is either equal in the both techniques and also some time error rate is lower as expected when we are using the bilateral filter. In figure 6.3 the black line shows the proposed algorithm with bilateral filter and also green line show the proposed algorithm without using the bilateral filter.

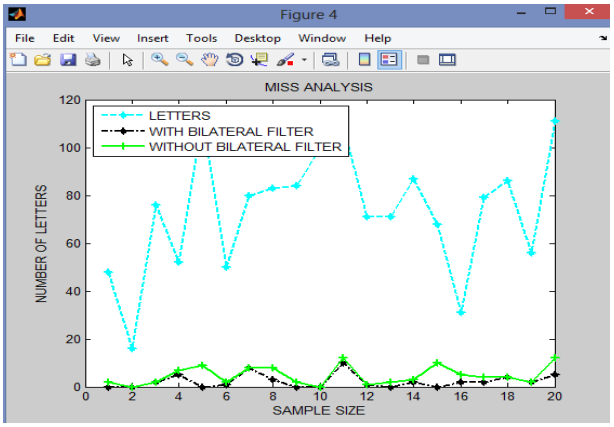

Figure 6.4 Miss Analysis

Figure 6.4is viewing the miss analysis of the proposed algorithm by considering the bilateral filter as well as without using bilateral filter. Figure 6.4 clearly express that the miss rate is either equal in the both techniques and also some time miss rate is lower as expected when we are using the bilateral filter. In figure 6.4 the black line shows the proposed algorithm with bilateral filter and also green line show the proposed algorithm without using the bilateral filter.

ISSN: 2278 – 1323

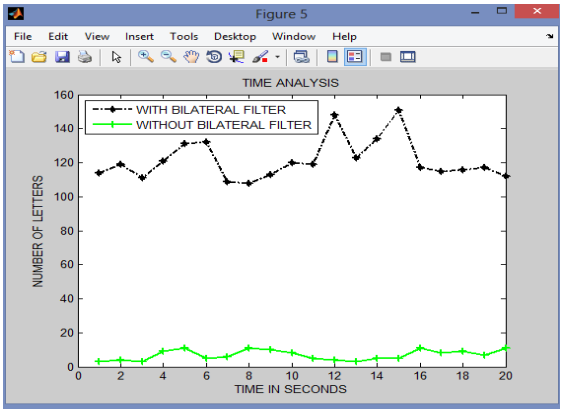

Figure 6.5 time analysis

Figure 6.3is viewing the time analysis of the proposed algorithm by considering the bilateral filter as well as without using bilateral filter. Figure 6.5 has shown that the proposed algorithm with bilateral is time consuming as it takes quite more time than without using the bilateral filter.

Conclusion

This research work has presented a new enhancement on English OCR technique by integrating it with well known bilateral filter. Various available techniques are studied to find the gaps among available techniques. But is found that the techniques which provide better results are slow in nature while fast techniques mostly provide inefficient results. It is found that the OCR techniques based on neural network provide more accurate results than other techniques.

But due to their speed problem we have developed a new technique which has used correlation to achieve the speed and better performance. It is found that proposed method has achieved up to 97.3% accuracy by using bilateral filter which was 94.33 % in Huang and Lin (2012) [16]. However training time is more in the proposed algorithm i.e. 167 seconds was 112 seconds in Huang and Lin (2012) [16]. And also recognition time is effective as it is 67 seconds as 136 in Huang and Lin (2012) [16].

The comparisons between proposed algorithm with bilateral filter and without it has shown that the bilateral filter can convert the input image in such a form so that the accuracy of the proposed algorithm become more consistent than without using the bilateral filter. However it has been shown by doing performance analysis that the proposed algorithm is time consuming when we use bilateral filter so this is the only limitation of the suggested algorithm.

In near future we will extend this work for integrating it will skew detection and correction techniques. In near future special characters are also introduced to improve the accuracy of the developed algorithm. Some more enhancements in matching is also done to decrease the computation time as well as increase the accuracy more.

Reference

[1] Dan Claudiu Cires,an and Ueli Meier and Luca Maria Gambardella and Jurgen Schmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", International Conference on Document Analysis and Recognition, IEEE, 2011.

[2] Georgios Vamvakas, Basilis Gatos, Stavros J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling" ,Pattern Recognition, Volume 43, Issue 8, August 2010.

[3] Shrey Dutta, Naveen Sankaran, Pramod Sankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N-Grams", IEEE, 2012.

[4] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, 2012.

[5] Yong-Qin Zhang, Yu Ding, Jin-Sheng Xiao, Jiaying Liu and Zongming Guo1, "Visibility enhancement using an image filtering approach", Zhang et al. EURASIP Journal on Advances in Signal Processing 2012.

[6] Sankaran, Naveen, and C. V. Jawahar. "Recognition of printed Devanagari text using BLSTM Neural Network." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[7] Vasudeva, Nisha, Hem Jyotsana Parashar, and Singh Vijendra. "Offline Character Recognition System Using Artificial Neural Network." International Journal of Machine Learning and Computing, 2012.

[8] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu. "A fast adaptive binarization method for complex scene images." In Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1889-1892. IEEE, 2012.

[9] Sumetphong , Chaivatna, and Supachai Tangwongsan. "An Optimal Approach Towards Recognizing Broken Thai Characters in OCR Systems." Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on. IEEE, 2012.

[10] AlSalman , AbdulMalik, et al. "A novel approach for Braille images segmentation." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, 2012.

[11] Mutholib, Abdul, Teddy Surya Gunawan, and Mira Kartiwi. "Design and implementation of automatic number plate recognition on android platform." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.

[12] Chi, Bingyu, and Youbin Chen. "Reduction of Bleed-through Effect in Images of Chinese Bank Items." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[13] Ramakrishnan, Kandan, and Evgeniy Bart. "Learning domain-specific feature descriptors for document images." Document Analysis Systems (DAS), 10th IAPR, 2012.

[14] Chattopadhyay, T., Ruchika Jain, and Bidyut B. Chaudhuri. "A novel low complexity TV video OCR system." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[15] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." Communications, Devices and Intelligent Systems (CODIS), 2012.

[16] Nan-Chi Huang and Huei-Yung Lin "A Multi-Stage Processing Technique for Character Recognition", The 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics 2012.

BIBLIOGRAPHY

Sukhpreet Singh received the B.Tech degree in Computer Science Engineering from the Baba Banda Singh Bahadur College of Engineering, Punjab Technical University Jalandhar(Punjab) in 2010, and M.Tech degree in Yadawindra College of Engineering University, Punjabi University, Patiala (Punjab) in 2013. His topic of interest is Optical CharacterRecognition (OCR).

.

ISSN: 2278 – 1323