Multilingual OCR (MOCR): An Approach to Classify Words to Languages

Mohammad Abu Obaida
Dept. of CSE, DUET
Gazipur, Bangladesh.

Md. Jakir Hossain
CSE, DUET
Bangladesh

Momotaz Begum
Lecturer
Dept. of CSE, DUET
Gazipur, Bangladesh

Md. Shahin Alam
CSE, DUET
Bangladesh

ABSTRACT
There are immense efforts to design a complete OCR for most of the world's leading languages, however, multilingual documents either of handwritten or of printed form. As a united attempt, Unicode based OCRs were studied mostly with some positive outcomes, despite the fact that a large character set slows down the recognition significantly. In this paper, we come out with a method to classify words to a language as the word segmentation is complete. For the purpose, we identified the characteristics of writings of several languages and utilized projecting method combined with some other feature extraction methods. In addition, this paper intends a modified statistical approach to correct the skewness before processing a segmented document. The proposed procedure, evaluated for a collection of both handwritten and printed documents, came with excellent outcomes in assigning words to languages.

General Terms
Pattern Recognition, Document Processing, Optical Character Recognition.

Keywords
OCR, Multilingual OCR, MOCR, Classification.

1. INTRODUCTION
The main focus of the efforts looking forward for a better recognition of characters in document, always aimed to excel and achieve a better precision in generating texts from the raw documents. Even an enormous of the researches dedicated to recognize the hand written documents that failed to accomplish 100% accuracy. However, multilingual OCR, a field of prospective is overlooked by and large, in spite of multilingual writing is harnessed worldwide especially in countries where several languages are officially approved.

Task of document skew correction, line and word segmentation are performed in MOCR in the similar fashion is done in a local or a particular OCR. Among the methods of skew correction Hough Transform [2] [3] [4] [5] [6], Projection Profile [7] [8] [9], Fourier transformation [10], Clustering [11], Cross Correlation [12] methods are established and utilized usually.

For the purpose of text line detection and segmentation a great diversity techniques for both printed handwritten documents has been studied and practiced widely. There are chiefly three fundamental categories that these text line detection and segmentation techniques fall in. Methods belonging in the first category employ the Hough transform [13] [14] [15] where, beginning from some points of the initial image, the lines that fit best to these points are extracted. The points considered in the Hough transform are typically either the gravity centers [13] or minima points [15] of the connected components. In [14] a block based Hough transform approach is applied taking the gravity centers of parts of connected components into account, which are called blocks. Methods belonging in the second category employ projections [1] [16] [17] [18] [22] [23] [24]. The methodology divides the document image into vertical strips as represented in [1] [17], and in these strips calculate the horizontal projections. The resultant projections are associated in order to extract the final text lines. In [16] [19], the histogram of the pixels' strength at each scan line is calculated. The produced bins are smoothed and the consequent valleys are identified. These valleys specify the space between the lines of the text. Lastly, methods of the third category use a kind of smearing [20] [21]. In [20], a fuzzy run length is used to segment lines. This measure is calculated for every pixel on the original image and describes how far one can see when standing at a pixel along horizontal direction. By this measure, a new grayscale image is created which is binarized and the lines of text are extracted from the new image.

Some other methods that do not belong in the previous categories are discussed in [25] [26] [27]. In [25], the text line extraction crisis is viewed from an Artificial Intelligence perspective. Shi et al. in [26], uses the Adaptive Local Connectivity Map. The method explained in [27], requires parameters to be set, according to the type of handwriting. Skew detection methods [28] [29] and baseline estimation methods [27], [28] are not flexible to capture the variation in handwriting. The method in [30], supposes that each connected component belongs to the same line.

For the purpose of word segmentation there exist two different tendencies. First of all, the connected components are calculated and the spaces between neighboring connected components are ascertained using a metric such as the Euclidean distance, the bounding box distance or the convex hull metric [19] [31] [32]. Then, a threshold is defined that is used to classify the calculated distances as either inter-word or inter-characters gaps. In the second group [33], the word segmentation crisis is assumed as a text line recognition task, adapted to the characteristics of segmentation. That is, at a particular point of a text line, it has to be found out whether the considered point belongs to a letter of a word, or to a space between words.

Most of the methods mentioned above, deal fruitfully with the separation of horizontally connected text lines which is a vital aspect towards word recognition; however, an ineffective inequity between inter-word and intra-word spaces. To be more exhaustive, none of the methods classify words to languages before character recognition thereby fail to identify foreign words accordingly convert them with junks. The reasons actuated us to present a methodology that harnesses complex projection profiles to facilitate words classification to languages immediately after line & word segmentation is complete therefore increasing the accuracy of character recognition.

2. MOCR: OPERATION PRINCIPLE

2.1 Projection Discrimination

In our MOCR we assumed two categories of languages.

Category 1: Languages that use English alphabets or modified (alike) form of English alphabets. Languages in this class are listed in Table 1. MoEn refers to this class of languafes.

Table 1. List of Languages that use English or alike alphabets

English	Latin	Italian	French
Spanish	Danish	Portuguese	German
Croatian	Turkish	Norwegian	Dutch
Polish	Latvian	Hungarian	Malay
Finnish	Romania	Vietnamese	Czech
Slovenian	Swedish	Indonesian	Irish
Hungarian	Filipino	Slovak	

Category 2(SASB Languages): Languages in this category use Stroke based alphabets; most of those belong to South Asia. Languages in this class are listed is Table 2. The phrase South Asian Stroke Based Languages is abbreviated to SASB Languages.

Table 2. List of Languages that use Stroke based alphabets

Bengali	Hindi	Nepali
Punjabi	Sanskrit	Marathi

The sentence "Freedom is the greatest achievement of any nation" is translated to languages listed in Table 1 & Table 2. Afterward, we calculated horizontal projection profiles for the sentences that are revealed in Table 3. According to the projection illustrated in the table we deduced, horizontal and vertical projection (illustrated in Figure 1) of languages in category 1& 2 profiles exhibit distinct and discriminating patterns that is listed in the Corollary 1 and Corollary 2.

Corollary 1: "The range where leading (upper amplitude) values belong in the profile, is in a distance than the class of values in previous larger range or the overall average amplitudes for stroke based languages. The range, in addition belongs to upper half of the projection for SASB languages"

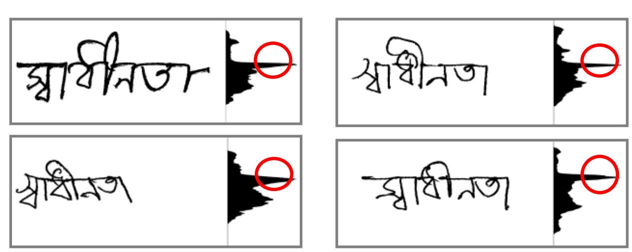
Corollary 2: "Intra-word spacing is absent (more or less) or negligible in South Asian Stroke based languages (Category 2)."

Here illustrated diverse types of Bengali handwritings of same word along with their horizontal projection in figure 2. That eventually evident Corollary 1 to be accurate.

2.2 Twofold Classifier

The twofold classifier, if employed, checks both the horizontal and vertical projections to confirm the classification of a word. In case of multilingual printed documents written in languages in category 1 &2; the method out performs as printed documents always restrain be intra-word spacing however multilingual handwritings as the slant is user dependent and character adjoin may cause the corollary 2 unviable. Therefore, printed documents validates both the corollaries however handwritings in some cases invalidate corollary 1. We can express in other words, corollary 1 to be writer independent but the corollary 2. In category-1 words as in Figure-1, if the writer does not link the characters or intra-word spacing exists in that case both the corollaries are viable.

Figure1.Vertical Projection for some multilingual text lines



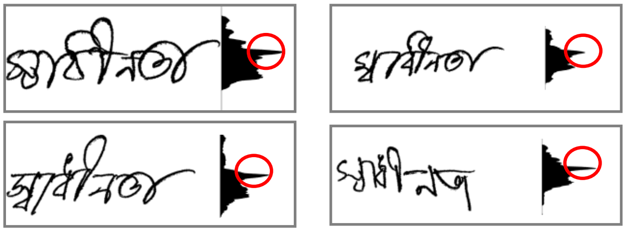
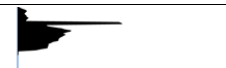
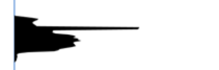
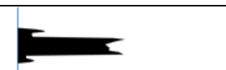


Figure 2: Horizontal projection of a word written in 8 different handwriting styles.

Table 3. List of Languages that use Stroke based alphabets

Language	Topic Sentence	Horizontal Projection
English	Freedom is the best achievement of any nation	
Bengali	স্বাধীনতা কোন জাতির সর্বশ্রেষ্ঠ অর্জন	
Hindi	आजादी के किसी भी राष्ट्र की सबसे बड़ी उपलब्धि है	
Marathi	फ्रीदोम इस थे ग्रेतेस्त अचीएवेमेंत ऑफ अन्य नातीव	
Punjabi	ਫ਼੍ਰੀਦੋਮ ਇਸ ਥੇ ਗ੍ਰੇਆਤੇਸ੍ਟ ਅਚਿਏਵੇਮੇਨ੍ਤ ਓਫ ਆਨੀ ਨਾਤਿਓਂ	
Sanskrit	फ़्रीदोम् इस थे ग्रेअतेस्त अचिएवेमेन्त ओफ़् अन्य नातिओन	
Nepali	फ्रीदोम इस द ग्रेअतेस्त अचिएवेमेन्त ओफ अन्य नेशन	
Croatian	Sloboa je najbolje ostvarenje bilo kojegnaroda	
Danish	Frihed er det bedste resultat af enhver nation	
Dutch	Vrijheid is de beste prestatie van elke natie	
Finnish	Vapaus on paras saavutus mille tahansa kansakunnalle	
French	La liberté est la meilleure réalisation de toute nation	
German	Freiheit ist die beste Leistung einer Nation	
Indonesian	Kebebasan adalah pencapaian terbaik dari negara manapun	
Italian	La libertà è il miglior risultato di ogni nazione	
Latin	Libertas optima facti alicujus gentis	
Norwegian	Frihet er den beste prestasjon av noen nasjon	
Portuguese	A liberdade é o melhor resultado de qualquer nação	
Spanish	La libertad es el mejor logro de una nación	
Swedish	Frihet är det bästa uppnåendet av någon nation	
Turkish	Özgürlük, herhangi bir ülkenin en iyi başarı	

3. OPERATION

A number of OCRs are available for most of the languages; we expect multilingual OCRs to combine features specific to particular OCRs, in other words, MOCR is a combined effort that accumulates several OCRs obligatory in order to convert a document. The words of different languages exhibit some specific patterns as the features (such as, horizontal or vertical profile) are extracted that is discussed in length at section 2.1.

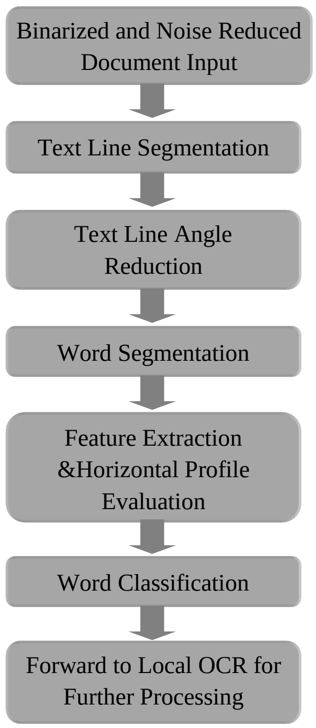


Figure 3: Process flow of MOCR Classification

The proposed method depicted in figure 3 assumes a Binarized (monochromic, gray scale) and noise reduced (by the means of adjusting the contrast and brightness or setting up the color level or discarding isolated clusters occupying few bits) document image as the input. The document, skew corrected in general, we apply line segmentation (discussed in length in the next subsection). After line segmentation is finished, we push surrounding borders along the grid points that the fencing to the text lines doesn't capture any text contents and white spaces are least. For handwritings, the line extracted might be slanted if words in the same line possess different obliquity. We optimize the skewness of the text line in the way obliquity is minimum for most of the words.

Line Segmentation: The text line segmentation section of proposed MOCR performs in the manner as statistical segmentation method presented in [1]; however a mere modification proposed. The process by default fix some values based on the document size that guides segmentation moving horizontally, vertically or diagonally, that eventually act like (*m x n*) grids over the document. Starting to segment lines, we start from left assumed grid points moving toward other end guided by the movement values as long as we don't cross any text content. We move diagonally or vertically in case we detect any obstacle of text contents; where diagonal movements are preferable to perpendicular. After an aberrant movement we again try to move flat always keeping in mind to exclude texts. The next line will stick to the same set of rules however must embrace a text line.

Word Segmentation: Word segmentation obliges that the document is already segmented in text lines. Besides, we also assume that for a text line, Connected Component (CC) belongs to a word, i.e. consecutive words are detached. Therefore, candidate word separators would lie at the gap between two successive CCs. We involve statistical projection profile method to segment words from lines. Projections (vertical) are very functional to identify words of SAS languages because of strokes. Therefore, only to decide, what the patterns and length of inter word spacing will be. We list all the spacing in a text line and discard extreme values; minute samples might be disregarded to facilitate the decision. Thus we remain with the spaces of a specific category, averaging the values we have a range of expected inter word spacing to consider.

Immediately after the words are segmented, the word classifier has the main role to play. For handwritings, words might be tilted, that we corrected (linearly straightened and upright) before applying to classifier; strokes there provide a useful parameter in correction with easily perceived statistical attribute. At this instant, we ascertain horizontal and vertical projection (if necessary) for the skew corrected word and apply both the corollaries (section 2.1) to classify the word. However, corollary 2 might not always be necessary; instead we employ corollary 2 that require vertical projection, only in the cases when there is complexity deciding the classification of a word.

The classified words finally fed to local OCR for further processing and recognition of characters and last of all, the character results are accumulated and arranged accordingly to form the final output.

4. MOCR PERFORMANCE EVOLUTION

We have collected 200 handwritten documents from peoples of ages 12 to 45 and evaluated the performance of our proposed methodology on a test set of 173 images. Total text lines in the test set were 2789 while the corresponding number for words was 16743; where 11045 words are stroke based Bengali, Hindi and remaining were written in English, Spanish in a mixed pattern e.g. figure 4, 6, 8. Among 173 test documents set there are two types of documents grouped with respect to intra-word spacing exists or not. Moreover, the texts were collected from guided, unguided and random writing sources. For guided texts (e.g. figure 4, 6) we provided the writers with texts and the direct them to write according to the guidelines. Unguided (figure 8) text writers were supplied with the texts but guidelines. Random texts are of diverse sources, they have their own set of writing and roughly no set of rules to follow.

Figure 5, 7, 9 delineate the line segmentation output, however the grids are imaginary formed with the movement values set by the method automatically. Figure 10, 11 depicts two sample final outputs of proposed word classifier; categorized words are set to different color levels.

To confirm the efficacy of our method we evaluated the performance of the modules specifically the classifier with respect to the actual data that is listed in Table 4.

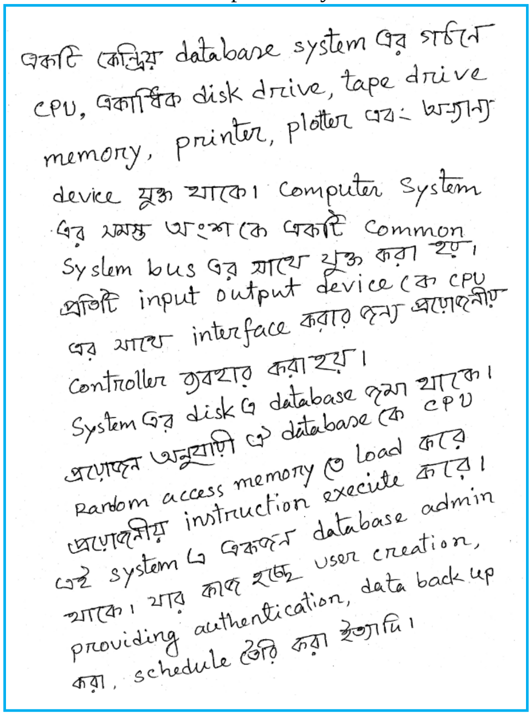


Figure 4: Raw Handwriting Image (Bengali+English)

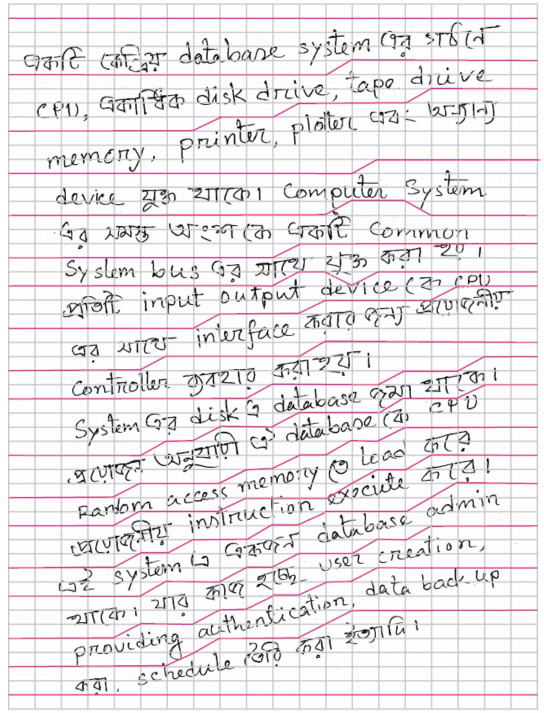


Figure 5: Line Segmentation Output

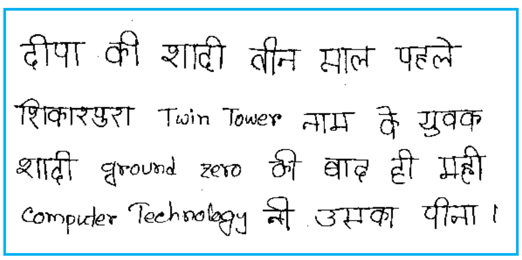


Figure 6: Raw Handwriting Image (Hindi+English)

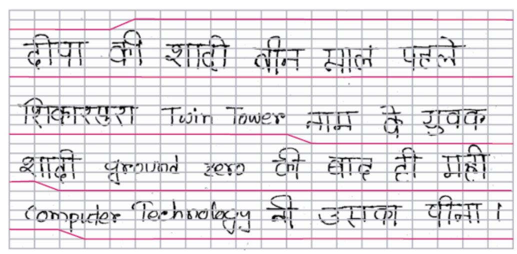


Figure 7: Line Segmentation Output

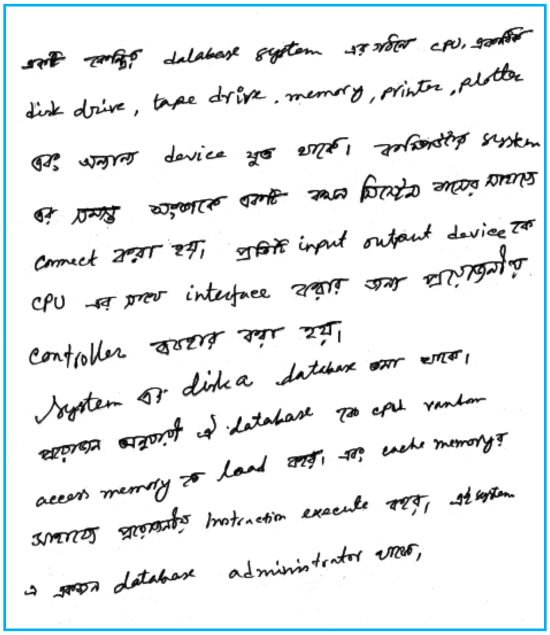


Figure 8: Raw Handwriting Image (Bengali+English)

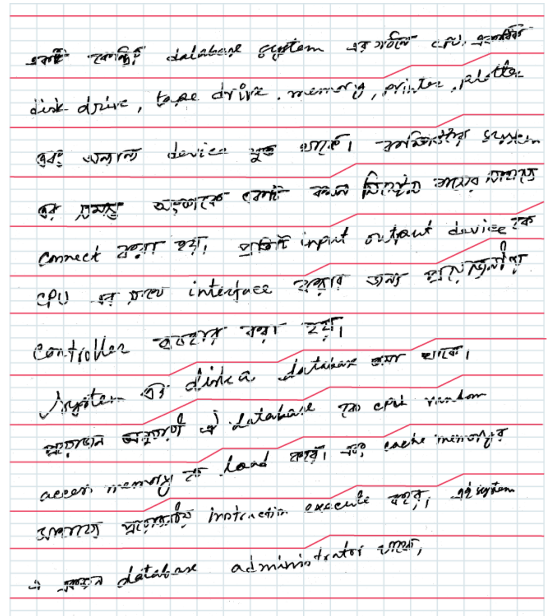


Figure 9: Line Segmentation Output

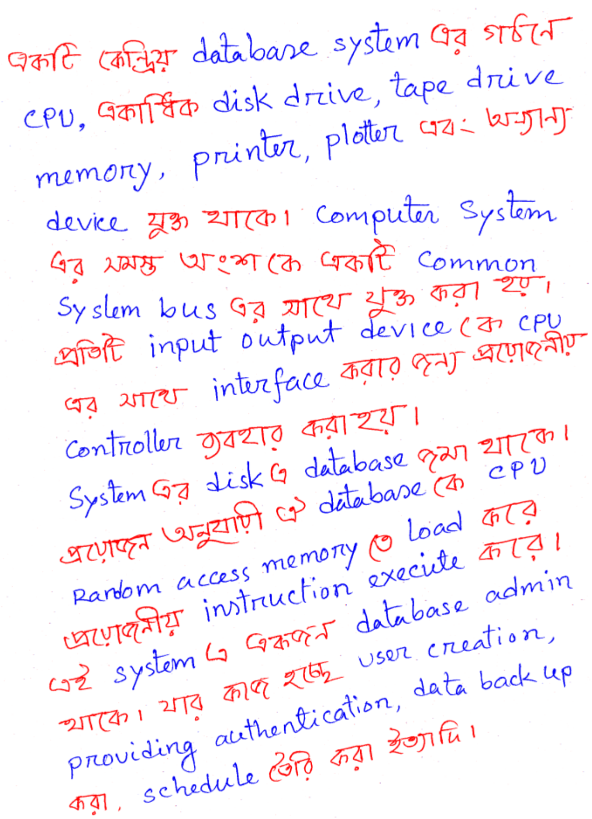


Figure 10: Final output after word classification (EN+BE)

Figure 11: Final output after word classification (EN+HI)

In the line segmentation phase, no complete grid vector is determined in real; traversing along the path looks like moving along a grid. The horizontal, vertical or diagonal movements for the segmentation of whole document constitute some fixed values that in terms define the accuracy of line segmentation. Inconsistent inter-word spacing and inter-word spacing cause troubles like same word parts are identified as two words and mixing of multilingual words part.

The number of errors encountered in the classification phase is due to the arbitrary and inconsistent writing style of the writers. In addition, some writer exploit different attribute in writing words of two languages alongside; that is considerable source of error in this phase.

Table 4. Performance Appraisal of Proposed Classifier

Segmentation and Classification				
	Documents	Line Seg. (det/actual)	Word Seg. (det/actual)	Classify Words (det/actual)
Guided Writings	64	1076/1088 (98.9%)	7513/7424 Error: 1.4%	SASB-4729/4800 (98.5%) EN-2784/2624 Error: 6.1%
Unguided Writings	57	923/1054 (97.3%)	6823/6612 Error: 3.2%	SASB-3856/4275 (90.2%) EN-2967/2337 Error: 27%
Random Writings	52	678/647 Error: 4.8%	2854/2707 Error: 5.4%	SASB-1912/1970 (97.1%) EN-942/737 Error: 27.8%
Total	173	2677/2789	17190/16743	SBSB-10497/11045 EN- 6693/5698

*SASB Language set consists of Bengali and Hindi texts
*Otherwise mentioned the percentage denotes accuracy

The accuracy percentage is ascertained by dividing detected values with actual values. We find the error percentage, in cases where number of detected words or lines exceeds actual value. For all types of writings a close scrutiny revealed that, MoEN (Category 1) words were categorized suitably with almost 99.86% accuracy and SASB (Category 2) words with 99.64% accuracy. Writings without stroke contribute as the main source of error. An overall of 12% error was committed in classification phase, which is writing dependent and higher for unguided and random writings.

5. CONCLUSION

In this paper, we present a classification methodology that categorizes multilingual texts either of handwritten or of printed document. Taking the previously published works on skew correction, on text line segmentation and on word segmentation into account the basic novelties of the proposed approach consist of (i) classifying words to languages before character recognition task is on track, (ii) an improved statistical methodology for the separation of vertically connected text lines and (iii) a hybrid word segmentation method recognizing inter-word and intra-word distances and judging vertical projection. The method evaluated with 173 handwritten documents, found to validate the propositions about classification with almost 91% accuracy except a bit more time consumption while processing large documents; that we expect to improve in our future works. In addition, we believe the MOCR methodology will be exploited widely as an obligatory module combining multiple local OCRs.

6. REFERENCES

[1] Manivannan Arivazhagan, H. Srinivasan, and S. N. Srihari: “A Statistical Approach to Handwritten Line Segmentation”, in Document Recognition and Retrieval XIV, Proceedings of SPIE, San Jose, CA, pp. 6500T-1-11, February 2007.

[2] Richard O. Duda, Peter E. Hart: “Use of the Hough Transform to deterct lines and curves in pictures”, Technical Note 36, AI Center, April 1971.

[3] Srihari, S.N. and V. Govindaraju: “Analysis of textual images using the Hough transforms”, Machine Vision Applications, 2: pp. 141-153, 1989. DOI: 10.1007/ BF01212455.

[4] D.S. Le, G.R. Thoma and H. Wechsler: “Automatic page orientation and skew angle detection for binary document images” Pattern Recognition, 27: pp. 1325-1344, 1994.

[5] U. Pal and B.B. Chaudhuri: “An improved document skew angle estimation technique”, Pattern Recognition Lett., 17: 899-904, 1996. DOI: 10.1016/0167-8655(96)00042-6

[6] Yu, B. and A.K. Jain: “A robust and fast skew detection algorithm for generic documents”, Pattern Recog., 29: 1599-1629, 1996. DOI: 10.1016/0031-3203(96)00020-9

[7] Mohammad Abu Obaida, Tanay Kumar Roy, Md. Abu Horaira and Md. Jakir Hossain: “Skew Correction Function of OCR: Stroke-Whitespace based Algorithmic Approach”, International Journal of Computer Applications (IJCA) 28(8):pp. 7-12, NY, USA, August 2011.

[8] H.S. Hou: “Digital Document Processing” Wisely New York, ISBN: 0471862479, 1983.

[9] Akiyama, T. and N. Hagita: “Automated entry system for printed documents”, Pattern Recognition, 23: 1141-1158, 1990. DOI: 10.1016/0031-3203(90)90112-X

[10] Omar, K., A. Ramli, R. Mahmod and M. Sulaiman: “Skew detection and correction of jawi images using gradient direction”, Journal of Technology, 37: 117-126, 2002.

[11] Hashizume, A., P.S. Yeh and A. Cosenfeld: “A method of detecting the orientation of aligned components”, Pattern Recognition Letters, 4: 125-132, 1986.

[12] Yan, H.: “Skew correction of document images using interline cross correlation”, Computer Vision Graph. Image Processing, 55: 538-543, 1993, DOI: 10.1006/cgip.1993.1041.

[13] Laurence Likforman-Sulem, Anahid Hanimyan, Claudie Faure: “A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents”, Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 774-777, 1995.

[14] G. Louloudis, K. Halatsis, B. Gatos, I. Pratikakis : “A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents”, 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006), La Baule, France, pp. 515-520, October 2006.

[15] Y. Pu and Z. Shi: “A Natural Learning Algorithm Based on Hough Transform for Text Lines Extraction in Handwritten Documents”, Proceedings of the 6th International Workshop on Frontiers in Handwriting Recognition, Taejon, Korea, pp. 637-646, 1998.

[16] Esra Ataer, Pinar Duygulu: “Retrieval of Ottoman Documents”, Proceedings of 8th ACM SIGMM International Workshop on Multimedia Information Retrieval , October 26-27, Santa Barbara, CA, USA, 2006.

[17] Elisabetta Bruzzone, Meri Cristina Coffetti: "An Algorithm for Extracting Cursive Text Lines", Proceedings of the Fifth International Conference on Document Analysis and Recognition, Bangalore, India, pp. 749, 1999.

[18] R. Manmatha, J. L. Rothfeder: "A Scale Space Approach for Automatically Segmenting Words from Historical Handwritten Documents", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.27, No.8, pp. 1212-1225, August 2005.

[19] U. Mahadevan, R. C. Nagabushnam: "Gap metrics for word separation in handwritten lines", Third International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 124-127, 1995.

[20] Z. Shi, V. Govindaraju: "Line Separation for Complex Document Images Using Fuzzy Runlength", First International Workshop on Document Image Analysis for Libraries, pp. 306, 2004.

[21] Wahl, F.M., Wong, K.Y., Casey R.G.: "Block Segmentation and Text Extraction in Mixed Text/Image Documents", Computer Graphics and Image Processing, 20, pp. 375-390, 1982.

[22] Berrin A. Yanikoglu, Peter A. Sandon: "Segmentation of off-line cursive handwriting using linear programming", Pattern Recognition 31, pp. 1825-1833, 1998.

[23] Ergina Kavallieratou, N. Dromazou, Nikos Fakotakis, George K. Kokkinakis: "An integrated system for handwritten document image processing", IJPRAI, International Journal of Pattern Recognition and Artificial Intelligence 17(4), pp. 101-120, 2003.

[24] C. Weliwitage, A. L. Harvey, B. Jennings: "Handwritten document offline text line segmentation", Digital Image Computing: Techniques and Applications, DICTA, pp. 184-187, December 2005.

[25] S. Nicolas, T. Paquet, L. Heutte: "Text Line Segmentation in Handwritten Document Using a Production System", Proceedings of the 9th IWFHR, Tokyo, Japan, pp. 245-250, 2004.

[26] Z. Shi, S. Setlur, and V. Govindaraju: "Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map", Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, pp. 794-798, 2005.

[27] Markus Feldbach and Klaus D. Tönnies: "Line detection and segmentation in historical church registers", Sixth International Conference on Document Analysis and Recognition, Seattle, USA, IEEE Computer Society , pp. 743-747, September 2001.

[28] Jisheng Liang, Ihsin T. Phillips and Robert M. Haralick: "A statistically based, highly accurate text line segmentation method", Proceedings 5th ICDAR, pp. 551-554, 1999.

[29] Douglas J. Kennard, William A. Barrett: "Separating lines of text in free-form handwritten historical documents", Second International Conferenceon Document Image Analysis for Libraries (DIAL), pp. 12-23, April 2006.

[30] Stéphane Nicolas, Thierry Paquet, Laurent Heutte: "Text line segmentation in handwritten document using a production system", Ninth International Work-shop on *Frontiers in Handwriting Recognition*, IWFHR-9 26-29, pp. 245-250, October 2004.

[31] U. V. Marti, H. Bunke: "Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition", *Sixth International Conference on Document Analysis and Recognition,* Seattle, WA, USA, pp. 159-163, 2001.

[32] G. Seni, E. Cohen: "External Word Segmentation of Off-line Handwritten Text Lines", *Pattern Recognition,* 27(1): 41-52, 1994.

[33] F. Luthy, T. Varga, H. Bunke: "Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation", *Ninth International Conference on Document Analysis and Recognition*, Curitiba, Brazil, pp. 8-12, 2007.

7. AUTHORS PROFILE

Mohammad Abu Obaida obtained Bachelor of Science in Engineering degree from Department of CSE, Dhaka University of Engineering & Technology (DUET), Gazipur, Bangladesh. At present wide-ranging research on networks, cryptography and pattern recognition are carried out by him. His key research interests include Cryptography, Networks and Web Security, Wireless networks, Software Architecture, Machine Vision, Artificial Intelligence, Protocol analysis and design. He has several research articles published, all in internationally distinguished journals.

Md. Jakir Hossain a obtained his B.Sc. in Engineering degree from Department of CSE, DUET, Bangladesh. Algorithms, Data Structures, Dynamic Programming, Patterns and Bioinformatics are the topics of major interest to him.

Momotaz Begum a graduate master's student, who obtained her B. SC. in Engineering degree from Department of CSE, DUET in 2008, serving as a faculty member to the department. Her fields of interest comprise of Pattern Recognition, Wireless networks and Network Security, Bioinformatics, Neural Networks.

Md. Shahin Alam achieved B. SC. in Engineering degree from Department of CSE, DUET in 2011.His fields of interest consist Networks and Pattern Recognition.