An Approach for Script Identification from Official Indic Scripts

Synopsis submitted

by

Sk Md Obaidullah

DOCTOR OF PHILOSOPHY (Engineering)

Department of Computer Science & Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata-700032
India

December, 2016

Synopsis

The dream of making a 'paperless world' will become a reality if an overwhelming volume of physical documents can be converted into its digital form. Researchers are working towards achieving this goal by developing several techniques for automatic processing of text document images. The initial step for developing such an auto processing system is digitization of the document files. Digitized text documents have several advantages, like indexing and sorting of large volumes of data, for efficient search operation and retrieval. Digitization of text documents can even ensure their preservation since digital documents will be protected from any kind of damage, degradation, the later being a common scenario in physical documents. In the past, researchers worldwide have exploited this possibility of digitization in an attempt to develop an image-to-alphanumeric text conversion system. Such a system is well renowned – popularly known as Optical Character Recognizer (OCR) [1]. The history of character recognition dates back to the year 1870 when the retina scanner system was invented by Carey [1], which was a photocell based image transmission system. In the late 1960's, soon after the invention of the digital computer, scientists realized the necessity of OCR for document processing system. As per record, the first commercialized OCR was developed by IBM to read the special font of IBM machines [1] [2]. Practitioners all around the world have, since then, been intrigued by this emerging field of research, which encompasses innumerable multi-faceted applications. Encapsulation of smart capabilities in the OCR system is such as, the ability to handle complex documents containing texts, color images and mathematical symbols, as well as, historical documents with degraded quality and noise, has been studied and is undoubtedly a booming field of study. The field is maturing day by day, by encapsulating smart capabilities in the system like, ability to handle complex documents which may contain text, graphics, mathematical symbols, historical documents with degraded quality and noise, color images etc. Smart ready to use commercial systems have been developed, whose applications include reading aids for blinds and automatic postal document sorter to name a few. Researchers belonging to the OCR community are now focusing on the development of efficient techniques for computerized document processing systems. However, in a multi-script country like India (having 11 scripts and 22 languages) [3], the prerequisite for these techniques forms an adequate knowledge of the particular script from which the language has been originated. Thus development of a script identification system

is essential in terms of the concerned research task. In our day to day life, we come across various multi-script documents such as postal documents filled up pre-printed application forms, railway reservation forms, etc. To handle such multi-script documents by OCR, script identification is an essential task before feeding the document images to language/script specific OCR. A block diagram of a multi-script document processing system is depicted in Figure 1. Initially, various multi-script document images are provided as input data. Then, basic pre-processing operations like noise removal, foreground-background separation, skew detection and correction, segmentation, etc. are performed. The next step performs script identification at page/block/line/word/character level where specific script type is produced as output. Then script dependent OCR is called from OCR bank and final textual information is generated.

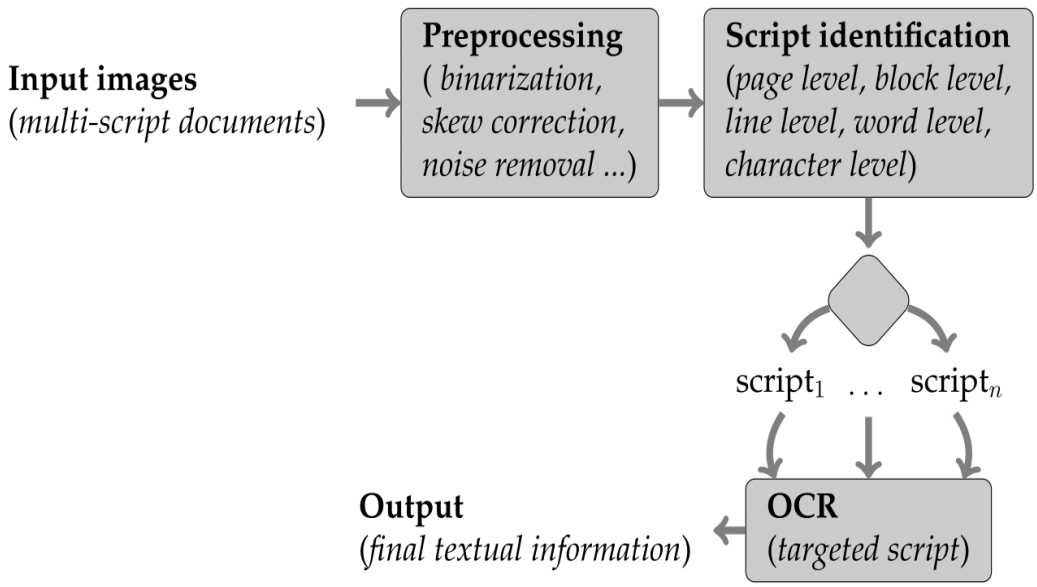

Figure 1 Block diagram of a multi-script document processing system showing different modules

The work presented in this thesis successfully handles various problems related to Indic script identification. To be more specific, in this endeavor, we have carried out the following tasks:

- Statistical analysis of different Indic languages and scripts with their demographic distribution
- Survey of Indic script identification techniques with their limitations

Synopsis

- Prepared handwritten script dataset for 11 Indic scripts which cover all the Indic languages. Prepared handwritten numeral scripts dataset from few of the most popular scripts.
- Printed and handwritten script identification from 11 official scripts with the analysis of bi-script, tri-script and multi-script performance
- Handwritten numeral script identification
- Study of the effect of document segmentation for script identification performance.

The work has been started with statistical analysis of different Indic languages and scripts with their demographic distribution. There are 23 official languages (including English) in India as per 8^{th} schedule of the Indian constitution [3]. These 23 languages are written using 11 different scripts, which means that, there are many languages which are written by a single script. Examples of such languages are Bangla (used to write Bengali, Assamese and Manipuri languages), Devanagari (used to write Hindi, Sanskrit, Nepali languages). Demographically speaking, Roman is the most popular script, followed by Devanagri and Bangla.

Some of the intrinsic properties of Indic scripts are as follows [4]:

- Scripts like Bangla and Devanagari contains a special topological property known as 'matra' or 'shirorekha'.
- Oriya and Malayalam scripts have components of a more circular shape than others.
- Considering Tamil script, most of the characters contain a 'T' like shape in their structure.
- Urdu script have maximum dot ('.') like small components. This script looks quite unlike other Indic scripts. Many characters of Urdu contain directional strokes with an orientation of about 75^0.
- There are many vertical, horizontal and slanting (about 45^0) strokes in Roman script.

- Kannada and Telugu scripts are quite similar, except a 'tick' like symbol present in Telugu script which is not there in Kannada. Similarly, Tamil and Malayalam characters are very much similar. Tamil and Malayalam characters have downward concavities and Kannada and Telugu characters have upward concavities.

The writing system of India follows an alphabetic writing system [5], which is divided into three main categories: abjad, abugida and true alphabetic. Urdu script, which has its origin in the Indo-European family falls under the abjad catagory. The Roman script follows true alphabetic system. The rest of the nine scripts belong to the Brahmic family of scripts, which fall under the abugida catagory. Brahmic family of scripts is divided into three classes: gupta, kadamba and grantha. All the eastern and northern Indian scripts are from gupta family. The four main south Indian scripts belong to the kadamba and grantha family.

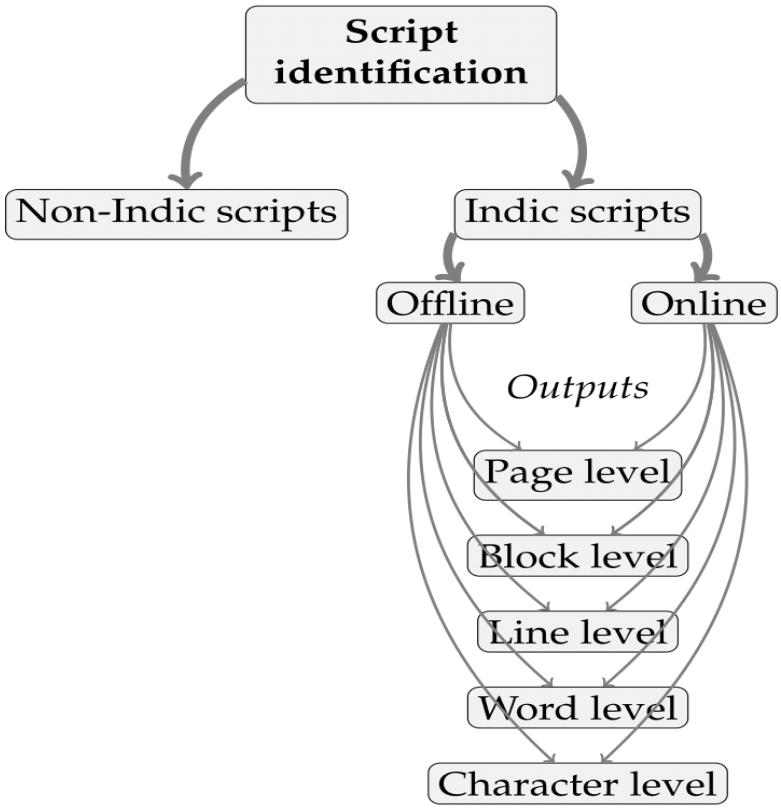

Figure 2 Catagorization of different script identification techniques

Synopsis

We have carried out a survey of handwritten script identification techniques. In this survey, we have divided all the works into several levels based on the type of documents considered as input. The diagram has been presented in Figure 2, which shows different offline script identification techniques at different levels: page, block, line, word and character. The level wise distribution of all these works has been discussed in this thesis. We found that, most of the works have been carried out at word and block level. Very few works are at page, line and character level. Motivated by this fact, we have prepared a 11-script page-level handwritten dataset and performed the script identification task [4]. At line-level, initially, we have proposed a script identification technique from eight official scripts. Later during multi-level script identification, we have performed script identification from all official Indic scripts at four major levels: page, block, line and word.

Availability of standard dataset for all official Indic scripts has been a real challenge for the script identification work. The issue of dataset development for script identification has been discussed. The state-of-the-art techniques on handwritten dataset development considering Indic scripts have been discussed and summarized. It can be deduced that, till date, handwritten Indic scripts dataset development has been restricted to a maximum of three scripts. This dataset is known as PBOK dataset [6], which consists of a total four scripts: Persian, Bangla, Oriya and Kannada, out of which last three are Indic scripts. From the global demographic distribution of different Indic scripts, we have found that, the Indic scripts considered in this thesis not only concerns of India but for researchers outside India too. We have discussed about the proposed dataset as a part of this thesis work. Although we have collected and used different printed/ handwritten datasets throughout this work, our main contribution is three handwritten datasets: (i) *PHDIndic_11* : a complete page-level handwritten dataset from 11 official Indic scripts [4] (ii) word-level printed dataset from 13 different languages and 11 official scripts [7] and (ii) *Numeral_db*: a handwritten numeral script dataset from four most popular Indic scripts [8]. We have presented PHDIndic_11 dataset, which consists of total 1458 pages from 11 different scripts written by 463 different writers and distributed over approximately 15010 lines and 124279 words. It is collected over duration of more than two years from different parts across the country. We have compared the proposed dataset with few of existing ones and illustrated the

effectiveness of the proposed one. We have proposed a word-level printed document image dataset from 13 different languages and 11 different scripts. The dataset consists of total 39k words, 3k words from each language. The script identification result on this dataset has been discussed. We have proposed the *Numeral_db* dataset, which is a handwritten numeral script dataset from four popular Indic scripts: Bangla, Devanagari, Roman and Urdu. This dataset consists of 5659 numeral strings written by 43 different writers. Additionally, we also have proposed benchmark results for script identification on these datasets.

Ghosh *et al.* [9] reported that, no universal feature exists which can effectively classify all the Indic scripts. Features are in general script/application dependent. Hence, for optimum performance, there might arise a need to combine different features through a heuristic feature selection approach. We have discussed about different feature extraction techniques. First, we have studied the shape and structural property of different scripts. Then based on our observation, we have computed different structural features: number of small components, presence of directional strokes, circularity, rectangularity and convexity of connected components, topological property etc. Here, one of our major contributions is optimizing the dimension of one of the topological feature, i.e. proposing a 1-dimensional fractal dimension (only one attribute is considered in this feature) [10] [11]. During the work of 'matra' and without 'matra' separation, we have compared the proposed 1-dimensional fractal dimension with two of the state-of-the-art techniques: canny edge detector and line transform. The effectiveness of the proposed features has been supported by experimental results. Fractal dimensions are used in handwritten script identification along with other features to obtain promising results. Another important contribution is the directional stroke based feature. Observing the presence of different directional strokes, we have defined four directional kernels, and feature values are computed applying different morphological operators. We have described different script independent features. Some state-of-the-art texture features namely: gray-level co-occurrence matrix, gabor filter bank, spatial energy, wavelet energy and radon transform have been studied. One of our contributions is optimizing the performance of wavelet features by making a feature fusion with radon transform i.e. we have proposed wavelet

radon transform or WRT [8]. Experimental results have shown the effectiveness of WRT in compared to normal wavelet transform. Another feature fusion based technique used is interpolated morphological transform or IMT [12]. It is a fusion of interpolation and morphological operations.

We have discussed the outcome of printed script identification. The present literature suggests that, most of the works have been carried out on printed documents [9]. This is due to less complexity of printed documents in comparison to handwritten one. Two different printed script identification problems have been addressed. In the first one, we have carried out page-level script identification from eleven official Indic scripts [13]. As no page-level printed dataset are available, we have conducted the experiment on our collected dataset. Mainly structural features or shape based features are used in this work. Performance of different classifiers is compared and random forest classifier is found to be the best performer with an average multi-script recognition accuracy of 98.99%, followed by LibLINEAR and MLP with 98.19% and 98.00 % respectively. This result can be considered as a benchmark on the dataset used in this work. In another problem, we have described the word-level script identification from eleven official Indic scripts (number of languages considered are thirteen) [7]. A dataset of volume 39k words, with equal distribution for each of the languages have been considered for the purpose of experiment. Three different features: spatial energy, wavelet energy and radon transform, three state-of-the-art classifiers: MLP, FURIA and random forests are used here. Two bi-script scenarios: (i) keeping Roman common with other languages (ii) keeping Devanagari common with other languages were considered. In scenario (i) we have received an average accuracy of 98.38% using MLP, while in scenario (ii) 99.24% average accuracy is obtained using the same classifier. During tri-script recognition (keeping both Roman and Devanagri common), we have received an average accuracy of 98.19% using MLP.

We have discussed about the handwritten script identification. The notable work reported are: (i) Block-level script identification from six official Indic scripts (ii) Line-level script identification from eight official Indic scripts (iii) Page-level script identification from eleven official Indic scripts (iv) Numeral script identification from four popular Indic

scripts (v) Script separation of 'matra' based scripts from scripts without 'matra' and (vi) Multi-level handwritten script identification from all official Indic script. Numeral script identification is a new direction of work in this field as it will help in different applications like: sorting of postal documents, arranging multi-lingual application forms or examination sheets based on the roll number written in candidates own scripts. We have conducted the experiment to separate scripts with 'matra' from scripts without 'matra' and used it as a precursor for script identification. Optimized 1-dimenstional fractal dimension has been used as the sole pertinent feature in this work. We have performed multi-level script identification from all the eleven official Indic scripts. In the literature, all the works had been carried out only at a single level. There is no theoretical or experimental justification available for choosing a particular level of work. So, here we have prepared a multi-level dataset, i.e. the same document has been considered at page, line, block and word level. Then, two different types of features: script dependent (structural) and script independent (global texture) have been applied at each level for script identification. So, in this work our objective is not only to study about the effect of segmentation on the performance of script identification but also to analyze which types of features are suitable at which level. Our observations are as follows: (i) line level data are more consistent irrespective of the features chosen. Block and page level data are comparatively similar and performance of word level data is very much feature dependent. (ii) Suitable feature combination has a remarkable effect on the overall performance of script identification (iii) Word level script identification is more challenging in terms of accuracy compared to page, line and block level identification. Hence, we have provided an experimental justification for choosing a particular level of work along with suitability of different features at different level of work. This is a new direction for future script identification work.

Finally we conclude that, the work presented in this thesis can be considered as an important step towards automation of document processing in the multi-script scenario. Here, the author proposes a pre-processing step (i.e. script identification) before supplying the document to script specific OCR. Some key issues of the script identification have been discussed with special emphasis on handwritten script identification. One of the important

outcomes of this thesis is presenting different frameworks and techniques for script identification, as well as, the development of benchmark handwritten datasets.

References

[1] J. Mantas, "An overview of Character Recognition Methodologies," *Pattern Recognition*, vol. 19, pp. 425–430, 1986.

[2] "OCR System: A Literature Survey." [Online]. Available: http://hdl.handle.net/10603/4166. last accessed 01.11.2016

[3] "Eigth Schedule," 2016. Available: http://mha.nic.in/hindi/sites/upload_files/mhahindi/files/pdf/Eighth_Schedule.pdf. last accessed 01.11.2016

[4] S. M. Obaidullah, C. Halder, K. C. Santosh, N. Das, and K. Roy, "PHDIndic_11: Page-level handwritten document image dataset of 11 official Indic scripts for script identification," *Multimedia Tools and Application*, communicated (under review), 2016.

[5] "Writting System of India." [Online]. Available: http://en.wikipedia.org/wiki/Writing_system. last accessed 01.11.2016

[6] A. Aleai, P. Nagabhushan, and U. Pal, "Dataset and Ground truth for Handwritten Text in Four Different Scripts," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 4, pp. 1253001 (25 pages), 2012.

[7] S. M. Obaidullah, K. C. Santosh, C. Halder, N. Das, and K. Roy, "Word-level thirteen official Indic languages database for script identification in multi-script documents," in *International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R 2016)*, 2016, (accepted).

[8] S. M. Obaidullah, C. Halder, N. Das, and K. Roy, "A New Dataset of Word-level Offline Handwritten Numeral Images from Four Official Indic Scripts and Its Benchmarking using Image Transform Fusion," *International Journal of Intelligent Engineering Informatics*, vol. 4, no. 1, pp. 1–20, 2016.

[9] D. Ghosh, T. Dube, and S. P. Shivprasad, "Script Recognition – A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010.

[10] S. M. Obaidullah, C. Goswami, K. C. Santosh, C. Halder, N. Das, and K. Roy, "Separating Indic scripts with `matra' -- a precursor to script identification in multi-script documents," in *IAPR International Conference on Computer Vision & Image Processing*, 2016, In Press.

[11] S. M. Obaidullah, C. Goswami, K. C. Santosh, C. Halder, N. Das, and K. Roy, "Separating Indic scripts with `matra' for effective handwritten script identification in multi-script documents," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 4, pp. 1753003 (17 pages), 2017.

[12] S. M. Obaidullah, C. Halder, N. Das, and R. Roy, "Bangla and Oriya Script Lines Identification from Handwritten Document Images in Tri-script Scenario," *International Journal of Service Science Management Engineering and Technology*, vol. 7, no. 1, pp. 43–60, 2016.

[13] S. M. Obaidullah, A. Mondal, N. Das, and K. Roy, "Script Identification from Printed Indian Document Images and Performance Evaluation Using Different Classifiers," *Applied Computational Intelligence and Soft Computing*, vol. 2014, pp. 12 pages, 2014.

List of Publications of the Author Related to the Thesis

Journal Publications

i. **Sk Md Obaidullah,** C Halder, N Das, K Roy, "Automatic Line-level Script Identification from Handwritten Document Images – A region-wise classification framework for Indian subcontinent", in Malaysian Journal of Computer Science (MJCS), 2016 **(accepted)**

ii. **Sk Md Obaidullah,** C Goswami, K C Santosh, C Halder, N Das, K Roy "Separating Indic scripts with 'matra' as a precursor for effective handwritten script identification in multi-script documents", in International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), World Scientific, vol. 31, no. 4, 1753003 (17 pages), 2017 **(in press)**

iii. **Sk Md Obaidullah,** C Halder, N Das, K Roy, "A New Dataset of Word-level Offline Handwritten Numeral Images from Four Official Indic Scripts and Its Benchmarking using Image Transform Fusion", in International Journal of Intelligent Engineering Informatics (IJIEI), Inderscience, vol. 4, no. 1, pp. 1-20, 2016

iv. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "Bangla and Oriya Script Lines Identification from Handwritten Document Images in Tri-script Scenario", in International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), IGI Global, vol. 7, issue 1, article 3, pp. 43-60, 2016

v. **Sk Md Obaidullah,** C Halder, N Das, K Roy, "Numeral Script Identification from Handwritten Document Images", in Procedia Computer Science Journal, Elsevier**,** vol. 54C (2015), pp. 585-594, 2015

vi. **Sk Md Obaidullah,** N Das, K Roy, "Convolution Based Technique for Indic Script Identification from Handwritten Document Images", in International Journal of Image, Graphics and Signal Processing, MECS Publisher, vol. 7 no. 5, pp. 49-57, 2015

vii. **Sk Md Obaidullah,** A Mondal, N Das, K Roy, "Script Identification from Printed Indian Document Images and Performance Evaluation using Different Classifiers", in Applied Computational Intelligence and Soft Computing, Hindawi Publishing Corporation, vol. 2014, Article ID 896128, 2014

viii. **Sk Md Obaidullah,** S K Das, K Roy, "A System for Handwritten Script Identification From Indian Document", in Journal of Pattern Recognition Research (JPRR), vol. 8, no. 1, 2013

ix. **Sk Md Obaidullah,** C Halder, K C Santosh, N Das, K Roy, "PHDIndic_11: Page-level handwritten document image dataset of 11 official Indic scripts for script identification", in Multimedia Tools and Applications (MTAP), Springer, 2016 **(under minor revision)**

x. **Sk Md Obaidullah,** C Halder, K C Santosh, N Das, K Roy, "Handwritten Indic Script Identification – A Survey", in Sadhana – Academy Proceedings of Engineering Sciences, Indian Academy of Science & Springer, 2016 **(revised version submitted)**

xi. **Sk Md Obaidullah,** K C Santosh, C Halder, N Das, K Roy "Automatic Indic script identification from handwritten documents – page, block, line and word-level approach", in Journal of Machine Learning and Cybernetics (JMLC), Springer, 2016 **(under review)**

Conference Publications

i. **Sk Md Obaidullah,** K C Santosh, C Halder, N Das, K Roy "Word-level thirteen official Indic languages database for script identification in multi-script documents", in International Conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R 2016), Karnataka, India, 2016 **(accepted)**

ii. **Sk Md Obaidullah,** C Goswami, K C Santosh, C Halder, N Das, K Roy, "Separating Indic scripts with `matra' -- a precursor to script identification in multi-script documents", in IAPR International Conference on Computer Vision & Image Processing (CVIP 2016), Roorkee, India, 2016

iii. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "PWDB_13: A Corpus of Word-level Printed Document Images for Thirteen Official Indic Scripts", in 4th International Conference on Frontiers on Intelligent Computing - Theory and Applications (FICTA 2015), Durgapur, India, 2015

iv. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "Visual Analytics based Technique for Handwritten Indic Script Identification – A Greedy Heuristic Feature Fusion Framework", in 4th International Conference on Frontiers on Intelligent Computing -

Theory and Applications (FICTA 2015), National Institute of Technology, Durgapur, India, 2015

v. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "An Approach for Automatic Indic Script Identification from Handwritten Document Images", 2nd Doctoral Symposium on Applied Computation and Security Systems (ACSS 2015), Calcutta, India, 2015

vi. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "Handwritten Indic Script Identification from Document Images - A Statistical Comparison of Different Attribute Selection Techniques in Multi-classifier Environment", in 2nd International Conference on Computer and Communication Technologies (IC3T 2015), Hyderabad, India, 2015

vii. **Sk Md Obaidullah,** C Halder, N Das, K Roy**,** "A Corpus of Word-level Offline Handwritten Numeral Images from Official Indic Scripts", in 2nd International Conference on Computer and Communication Technologies (IC3T 2015), Hyderabad, India, 2015

viii. **Sk Md Obaidullah,** C Halder, N Das, K Roy, "Indic Script Identification from Handwritten Document Images – An Unconstrained Block-level Approach", in 2nd International Conference on Recent Trends in Information Systems (ReTIS 2015), Jadavpur, Kolkata, India, 2015

ix. **Sk Md Obaidullah,** R Karim, S Shaikh, C Halder, N Das, K Roy**,** "Transform Based Approach for Indic Script Identification from Handwritten Document Images", in 3rd International Conference on Signal Processing, Communications and Networking (ICSCN 2015), Chennai, India, pp. 1-7, 2015

x. **Sk Md Obaidullah,** Z Rahaman, N Das, K Roy, "Development of Document Image Database for Offline Handwritten Indic Script Identification - A State-of-the-art", in International Conference on Pattern Recognition and Multimedia Signal Processing (ICPRMSP 2015), Tamilnadu, 2015

xi. **Sk Md Obaidullah,** N Das, K Roy**,** "Gabor Filter Based Technique for Offline Indic Script Identification from Handwritten Document Images", in International Conference on Devices, Circuits and Communication (ICDCCom 2014), Mesra, Ranchi, India, 2014.

xii. **Sk Md Obaidullah,** N Das, K Roy, "Offline Handwritten Script Identification from Eastern Indian Document Images using Logistic Model Tree", in 2014 International Conference on Intelligent Computing, Communication and Devices (ICCD 2014), Bhubaneswar, India, 2014

xiii. **Sk Md Obaidullah,** A Mondal, K Roy, "Structural Feature Based Approach for Script Identification from Printed Indian Document", in International Conference on Signal Processing and Integrated Network (SPIN 2014), Noida, India, 2014.

xiv. **Sk Md Obaidullah,** K Roy, N Das, "Comparison of Different Classifiers for Script Identification from Handwritten Document", in International Conference on Signal Processing, Computing and Control (ISPCC 2013), Shimla, India, 2013

xv. K Roy, S K Das, **Sk Md Obaidullah**, "Script Identification from Handwritten Document", in IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIGP 2011), Hubli, Karnataka, India, 2011