A Study on Script Identification from Document Images with Added Emphasis on Tamil Language

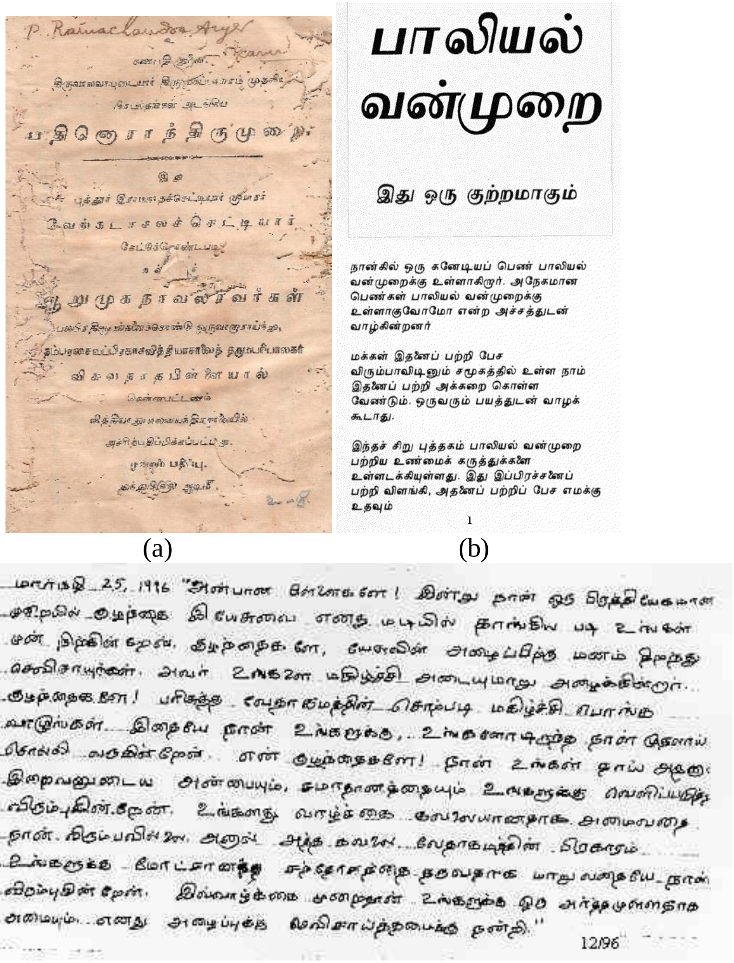
Faustina Joan S P^1, BharathiPriyanka L Y^2, Catherine Mary Philips^3
^{1}Assistant Professor, 23 UG Student
^{123}Department of Computer Science
Stella Maris College, Chennai

Abstract- Language is an integral part of everyday human communication and a powerful tool of expression through speech and writing. With increased technological advancements, there comes a need to enable computers, recognize languages and their scripts automatically to help many language dependent applications like speech recognition, document indexing, optical character recognition(OCR) and so on. Identifying the written language scripts from document images is a happening and a very challenging research area due to the availability of different languages, their writing systems and acquisition mediums. Research on English script identification is enormous and considerable research is being done on Indic languages and other foreign languages. Looking at the current literature, works specific to Tamil language script identification is sparse and there are few works that includes Tamil script as part of multi-script identification. This paper aims to give an outlook on various script writing systems, their identification, works related to the domain with special focus on Tamil script along with a generic prototype of Tamil script identification.

Keywords- pattern recognition, OCR, script identification, Tamil language, image processing

I. INTRODUCTION

Language is a human ability and a system of communication comprising of words articulated in a structured manner which can be either spoken or written.With today's technology, language also finds its placevisually in the form of textwhich are captured in images. According to Jung et al. [1] image content can be categorized as perceptual content and semantic content. When perceptual content concentrates on the visual perception and its attributes, semantic content derives meanings from the image through objects, text and events captured in them. Text is thereby an important feature that can be extracted from an image to understand its context. Images with text can be classified into born-digital, scene-text and document images. Born-digital are digitally created images with text whereas scene-text images have text naturally present in the scenario. Document images include text from scanned documents such as graphical documents, historical documents and papers. Figure 1 shows some examples of document images.



P. Ramachandra Aiyar
Karur

கணபதி துணை.

திருவாலவாயுடையார் திருமுகப்பாசுரம் முதலிய

பிரபந்தங்கள் அடங்கிய

பதினொராந் திருமுறை.

இது

புத்தூர் இராமநாதச்செட்டியார் குமாரர்

வேங்கடாசலச்செட்டியார்

கேட்டுக்கொண்டபடி

நல்லூர்

ஆறுமுகநாவலரவர்கள்

பலபிரதிரூபங்களைக்கொண்டு ஒருவாறாராய்ந்து,

தம்பரசைவப்பிரகாசவித்தியாசாலைத் தருமபரிபாலகர்

விசுவநாதபிள்ளையால்

சென்னபட்டணம்

வித்தியாநுபாலனயந்திரசாலையில்

அச்சிற்பதிப்பிக்கப்பட்டது.

மூன்றாம் பதிப்பு.

சுக்கில௵ ஆடிமீ.

(a)

பாலியல் வன்முறை

இது ஒரு குற்றமாகும்

நான்கில் ஒரு கனேடியப் பெண் பாலியல் வன்முறைக்கு உள்ளாகிறார். அநேகமான பெண்கள் பாலியல் வன்முறைக்கு உள்ளாகுவோமோ என்ற அச்சத்துடன் வாழ்கின்றனர்

மக்கள் இதனைப் பற்றி பேச விரும்பாவிடினும் சமூகத்தில் உள்ள நாம் இதனைப் பற்றி அக்கறை கொள்ள வேண்டும். ஒருவரும் பயத்துடன் வாழக் கூடாது.

இந்தச் சிறு புத்தகம் பாலியல் வன்முறை பற்றிய உண்மைக் கருத்துக்களை உள்ளடக்கியுள்ளது. இது இப்பிரச்சனைப் பற்றி விளங்கி, அதனைப் பற்றிப் பேச எமக்கு உதவும்

1

(b)

மார்கழி 25, 1996 "அன்பான செல்வங்களே! இன்று நான் ஒரு பிரத்தியேகமான முறையில் குழந்தை இயேசுவை எனது மடியில் தாங்கிய படி உங்கள் முன் நிற்கின்றேன். குழந்தைகளே, யேசுவின் அழைப்பிற்கு மனம் திறந்து செவிசாயுங்கள். அவர் உங்களை மகிழ்ச்சி அடையுமாறு அழைக்கின்றார். குழந்தைகளே! பரிசுத்த வேதாகமத்தின் சொற்படி மகிழ்ச்சி பொங்க வாழுங்கள். இதையே நான் உங்களுக்கு, உங்களோடிருந்த நாள் முதலாய் சொல்லி வருகின்றேன். என் குழந்தைகளே! நான் உங்கள் தாய் அருளு: இறைவனுடைய அன்பையும், சமாதானத்தையும் உங்களுக்கு வெளிப்படுத்த விரும்புகின்றேன். உங்களது வாழ்க்கை கொடையானதாக அமைவதை நான் விரும்புகின்றேன். அதை அந்த கடவுள் வேதாகமத்தின் பிரகாரம் உங்களுக்கு மோட்சானந்த சந்தோசத்தை தருவதாக யாது அதையே நான் விரும்புகின்றேன். இவ்வாழ்க்கை முறைதான் உங்களுக்கு ஒரு அர்த்தமுள்ளதாக அமையும். எனது அழைப்புக்கு செவிசாய்த்தமைக்கு நன்றி." 12/96

(c)

Fig.1: Examples of Scanned Document Images in Tamil
(a) Historical Documents (b) Page Documents (c) Handwritten Documents

Every document image has numerous lines of text with either single language script or multiple scripts. Identifying the language and processing it enables many applications such as automatic indexing, text recognition, document analysis, content-based retrieval and tourist translators. Therefore, script identification is crucial and requires utmost accuracy to retain the script and the meaning it conveys. This paper deals with document images,emphasizing on script identification, works related to it with special focus on Tamil script identification. In the context of the paper, script and language are considered synonymous though a script type can contain one or more languages under them. The following section briefs about the types of script writing systems.

II. TYPES OF SCRIPT WRITING SYSTEMS

The world's writing systems can be divided into 6 types namely logographic, alphabetic, syllabic, featural, Abjads, Abugidas systems [2]. The logographic system is used in Chinese and Korean languages which has full word representations mostly.Languages like English, Latin, Russian use the alphabetic system where the language is constructed on alphabets. A syllable represented every symbol in the syllabic system as in the Japanese language which is also a logographic language. The Korean language is an example of featural writing system which is built on phonetics.The Abjads system consists of Hebrew and Arabic in which consonant sounds make up every symbol. Almost all South-east Asian and Indian languages whose source is the Brahmic script follow the Abugidas writing system.It is based on the concept of consonant-vowel unit and stands between the alphabetic and syllabic writing systems. Tamil, Telugu, Malayalam, Kannada, Bangla, Manipuri, Oriya are languages that follow Abugidas.Figure 2 lists the different script writing systems with their languages.

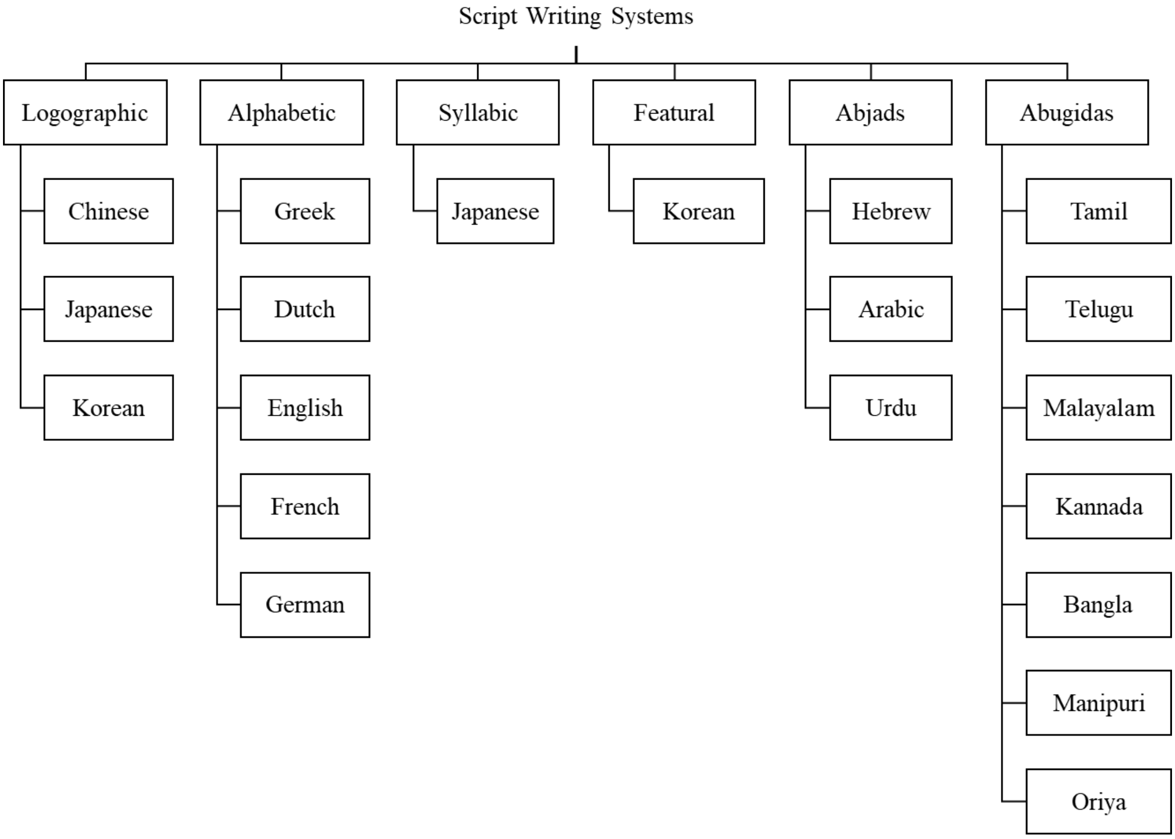
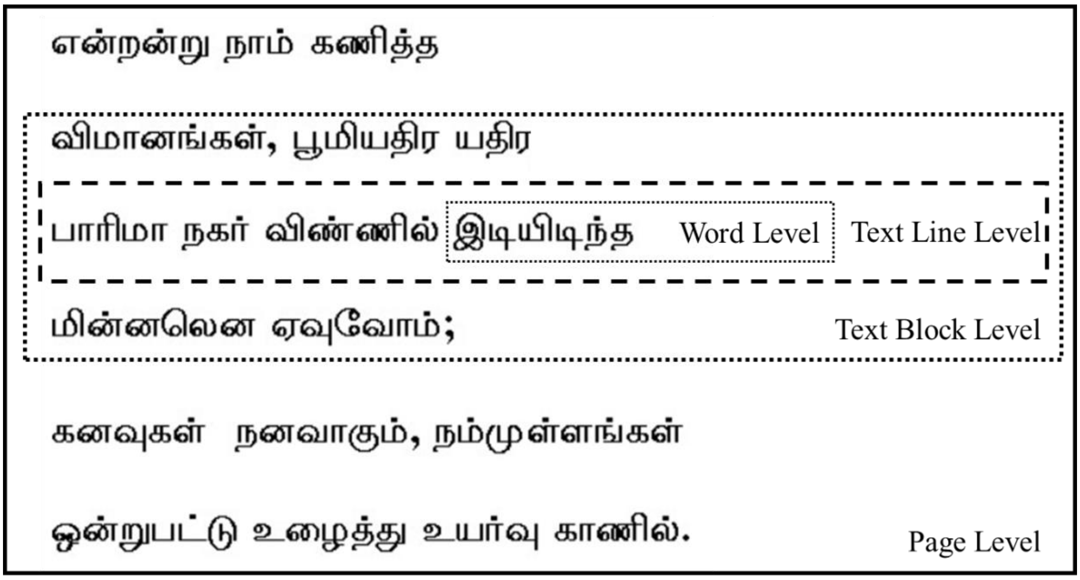


Fig.2: Classification of Script Writing Systems

The biggest challenge of script identification lies on the numerous languages that come under the script writing system with identical characters of same stroke and shape. Some languages can be differentiated by discrete alphabets whereas some languages follow a unit based pattern. Also the medium of document images has to be considered during identification. So a single identification methods will not suffice every language and medium. The next section describes about script identification.

III. SCRIPT IDENTIFICATION

Document images can be categorized as printed, handwritten and hybrid based on how the script is sourced and contented. There is concern regarding the mode of acquiring these images as it influences the overall image quality and the script clarity. Taking these into consideration, identification can be classified into different levels namely text-block, text line, word, and character levels [2] as shown in Figure 3. It involves identifying the text in the documents followed by the script. The following sections introduces the various identification methods.



என்றன்று நாம் கணித்த
விமானங்கள், பூமியதிர யதிர
பாரிமா நகர் விண்ணில் இடியிடிந்த
மின்னலென ஏவுவோம்;
கனவுகள் நனவாகும், நம்முள்ளங்கள்
ஒன்றுபட்டு உழைத்து உயர்வு காணில்.

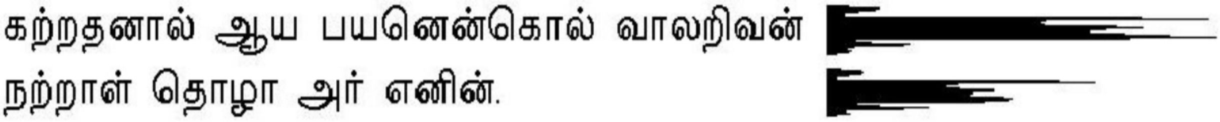
Fig.3: Script Identification Levels

A. Text-block Identification

This identification is also known as page level or document level identification. Every page or document is divided into many text-blocks which in turn has many lines and words present in them. There are many methods that have been explored under this category. Hassan et al. [3] identified the script in multiple levels using texture and shape information starting from page-level identification. Singh et al. [4] used Gray Level Co-occurrence Matrix (GLCM) to extract texture features at page-level to identify the script. The methods are not restricted to only one level. For example, GLCM can also be used in text line and character identification.

B. Text Line Identification

The next level of identification starts at the text line level. A text line is a combination of words and spaces. Researchers either start with the identification of these text lines directly or identify them after the page level. Horizontal projection profile depicted in Figure 4, is one of the most commonly used methods for text-line identification. Projection profile is the running sum of the pixels in either horizontal or vertical direction.



கற்றதனால் ஆய பயனென்கொல் வாலறிவன்
நற்றாள் தொழா அர் எனின்.

Fig.4: Horizontal Projection Profile

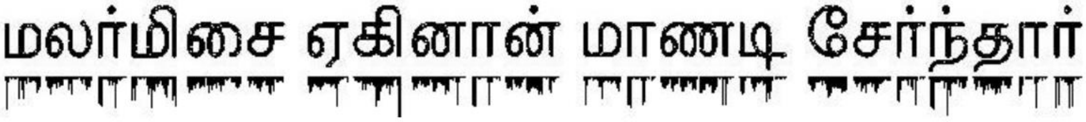
Busch et al. [5] used projection profiles in their work to identify script using textures in the text line level.Pal et al. [6] identified multiple scripts from Indian documents using valley and peak information of projection profiles.

C. Word Identification

Pati et al. [7] used Gabor and DCT features to identify words in a script and implemented it to identify the script to which the word belonged to. Singh et al. [8] used a combination of shape and texture based features to identify multiple scripts. At word level, features are extracted from them and scripts are identified through image processing techniques or machine learning.

D. Character Identification

Earlier research works used vertical projection profile as shown in Figure 5 to segment characters.



மலர்மிசை ஏகினான் மாணடி சேர்ந்தார்

Fig.5: Vertical Projection Profile

Character extraction is a very important step when the script needs to be recognized using Optical Character Recognition (OCR). Many researchers like Pal & Sarkar [9] who work on script recognition concentrate on character level identification. The following section lists some of the recent research works on script identification.

IV. RECENT RESEARCH WORKS AND TAMIL SCRIPT IDENTIFICATION

Ubul et al. [2], Bashir et al. [11], Sahare&Dhok [12] have recently come up with an extensive survey on script identification. Script identification is either restricted to English or scripts belonging to a specific writing system. Kavitha et al. [13] found out the possibilities of using the spatial relationships of intersection, end and junction points of connected components to find out the component structure. These points helped in identifying historical documents in Tamil, English and few other Indic languages. Rumma [14] used Radon and GLCM features along with machine learning to identify 6 languages including Tamil.Research on multi-script Indic language identification is on the high though works concentrating only on Tamil language identification is not significant. Also no paper yet has come up with 100% accuracy of identifying a script. A generic prototype of Tamil script identification is given in Figure 6.

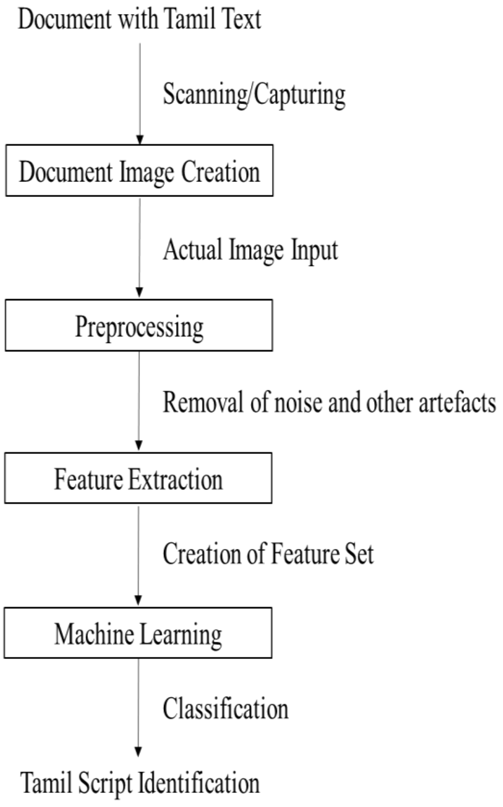
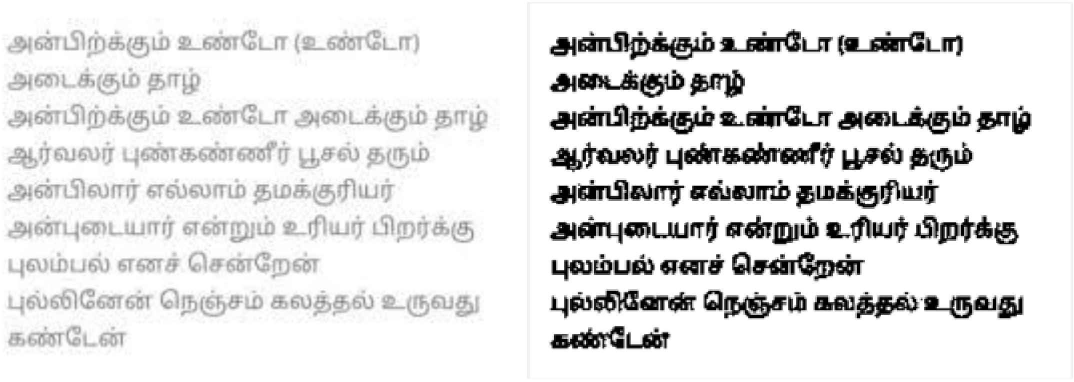


Fig.6: Generic Prototype of Tamil Script Identification

First, the document with Tamil text has to be scanned or captured by imaging devices like camera so that it is converted into a document image.This image may contain noise and artefacts like blur and distortion. So, to remove them the image needs to be preprocessed using image processing techniques. For example, in Figure 7 the image is preprocessed using binarization technique to remove noise and increase the contrast of the input image.


அன்பிற்க்கும் உண்டோ (உண்டோ)
அடைக்கும் தாழ்
அன்பிற்க்கும் உண்டோ அடைக்கும் தாழ்
ஆர்வலர் புண்கண்ணீர் பூசல் தரும்
அன்பிலார் எல்லாம் தமக்குரியர்
அன்புடையார் என்றும் உரியர் பிறர்க்கு
புலம்பல் எனச் சென்றேன்
புல்லினேன் நெஞ்சம் கலத்தல் உருவது
கண்டேன்

அன்பிற்க்கும் உண்டோ (உண்டோ)
அடைக்கும் தாழ்
அன்பிற்க்கும் உண்டோ அடைக்கும் தாழ்
ஆர்வலர் புண்கண்ணீர் பூசல் தரும்
அன்பிலார் எல்லாம் தமக்குரியர்
அன்புடையார் என்றும் உரியர் பிறர்க்கு
புலம்பல் எனச் சென்றேன்
புல்லினேன் நெஞ்சம் கலத்தல் உருவது
கண்டேன்

(a) (b)

Fig.7: (a) Input Image before Preprocessing (b) Preprocessed Image

After preprocessing, features to identify the text in the document has to be extracted so that a feature set is formed. The features can be based on the structure, shape and component at page, text block, text line, word or character level. Using the extracted feature set, machine learning techniques will classify the text to be a valid Tamil script or not. Based on the classification results, Tamil script will be identified. The prototype can be extended to recognize the identified script. The next section concludes the paper with future scope.

V. CONCLUSION AND FUTURE WORK

This paper aims to give an understanding about the domain of script identification with an introduction to the generic prototype of Tamil language script identification. Though there is considerable research for English and multi-script identification, it is known that works pertaining to Tamil scripts is few. If 100% accuracy can be reached in identifying the script, many applications can be successful built for the society. To come up with such an application will be the goal wherein our future work will first concentrate on writing appropriate preprocessing methods and finding unique features that will help in identifying the script. Through the paper, it is also known that this interesting domain has room for improvements and needs further research ideas to achieve full accuracy.

VI. REFERENCES

[1]. Jung K, Kim K, Jain AK (2004) Text Information Extraction in Images and Video: A Survey. Pattern Recogn 37(5):977–997.

[2]. Ubul, K., Tursun, G., Aysa, A., Impedovo, D., Pirlo, G., &Yibulayin, T. (2017). Script Identification of Multi-Script Documents: A Survey. IEEE Access, 5, 6546-6559.

[3]. Hassan, E., Garg, R., Chaudhury, S., & Gopal, M. (2011). Script based text identification: a multi-level architecture. In Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data (p. 11). ACM.

[4]. Singh, Pawan Kumar et al. (2015) Page-level script identification from multi-script handwritten documents. Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT) (2015): 1-6.

[5]. Busch, A., Boles, W. W., &Sridharan, S. (2005). Texture for script identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11), 1720-1732.

[6]. Pal, U., Sinha, S., &Chaudhuri, B. B. (2003). Multi-script line identification from Indian documents. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003 (p. 880). IEEE.

[7]. Pati, P. B., &Ramakrishnan, A. G. (2008). Word level multi-script identification. Pattern Recognition Letters, 29(9), 1218-1229.

[8]. Singh, P. K., Mondal, A., Bhowmik, S., Sarkar, R., &Nasipuri, M. (2015). Word-level script identification from handwritten multi-script documents. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, Springer, Cham, 551-558.

[9]. Pal, U., & Sarkar, A. (2003). Recognition of printed Urdu script. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003 (p. 1183). IEEE.

[10]. Obaidullah, S. M., Goswami, C., Santosh, K. C., Halder, C., Das, N., & Roy, K. (2017). Separating Indic Scripts with 'matra'—A Precursor to Script Identification in Multi-script Documents. In Proceedings of International Conference on Computer Vision and Image Processing (pp. 205-214). Springer, Singapore.

[11]. Bashir, R., Quadri, S. M. K., &Giri, K. J. (2018). Script identification: a review. International Journal of Information Technology, 1-15.

[12]. Sahare, P., &Dhok, S. B. (2017). Script identification algorithms: a survey. International Journal of Multimedia Information Retrieval, 6(3), 211-232.

[13]. Kavitha, S., Shivakumara, P., Kumar, G. H., & Tan, C. L. (2015). A robust script identification system for historical Indian document images. Malaysian Journal of Computer Science, 28(4), 283-300.

[14]. Rumma, S. S. (2018) Word-wise South Indian Script Identification using GLCM and Radon Features. International Journal on Future Revolution in Computer Science & Communication Engineering, 4(2), 476-478.

[15]. Obaidullah, S. M., Mondal, A., & Roy, K. (2014). Structural feature based approach for script identification from printed Indian document. In Signal Processing and Integrated Networks (SPIN), 2014 International Conference on (pp. 120-124). IEEE.