IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received January 26, 2017, accepted February 16, 2017, date of publication March 30, 2017, date of current version May 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2689159

Script Identification of Multi-Script Documents: A Survey

KURBAN UBUL[1], (Member, IEEE), GULZIRA TURSUN[1], ALIMJAN AYSA[2], (Member, IEEE), DONATO IMPEDOVO[3], (Member, IEEE), GIUSEPPE PIRLO[3], (Senior Member, IEEE), AND TUERGEN YIBULAYIN[4], (Member, IEEE)

[1]School of Information Science and Engineering, Xinjiang University, Ürümqi 830046, China
[2]Network and Information Technology Center, Xinjiang University, Ürümqi 830046, China
[3]Dipartimento di Informatica, University of Bari, 70126 Bari, Italy
[4]Xinjiang Laboratory of Multi-language Information Technology, Ürümqi, Xinjiang 830046, China

Corresponding author: Donato Impedovo (donato.impedovo@uniba.it)

This work was supported in part by the National Natural Science Foundation of China under Grant 61163028, Grant 61363064, Grant 61563052, in part by the Special Training Plan Project of Xinjiang Uyghur Autonomous Region's Minority Science and Technological Talents under Grant 201323121, in part by the College Scientific Research Plan Project of Xinjiang Uyghur Autonomous Region under Grant XJEDU2013I11, in part by the Open Projects Program of National Laboratory of Pattern Recognition under Grant 201306321, and in part by the Italian Ministry of Education, University and Research within the PRIN2015-HAND Project under Grant H96J16000820001.

ABSTRACT In recent years, with the widespread of Internet and digitized processing of multi-script documents worldwide, script identification techniques have become more important in the pattern recognition field. Script identification concerns methods for identifying different scripts in multi-lingual, multi-script documents. This paper presents a comprehensive overview on research activities in the field and focuses on the most valuable results obtained so far. The most vital processes in script identification are addressed in detail: identification and discriminating methods, features extraction (local and global), and classification. Different kinds of approaches have been developed and promising results have been achieved. This paper reports SoA performance results. This paper reports methods concerning handwritten, printed, and hybrid document processing. More research is necessary to meet the performance levels essential for everyday applications.

INDEX TERMS Handwriting recognition, optical character recognition (OCR), character recognition, multi-script documents, script identification.

I. INTRODUCTION

Our reliance on the digital world is continuously increasing with the rapid developments in information and technology in all aspects of our lives. In administrative and office environments, the development of Optical Character Recognition (OCR) systems have lightened the possibility of creating paperless solutions. Unfortunately, although the field of OCR has been one of the oldest and more investigated research fields, still today OCR systems are just specialized in one particular script. In order to overcome this limitation, script identification has been used. In fact, in a multi-script document image, script identification is necessary to find text portions written in the same script, so that script-specific OCR system can be applied. Hence, script identification system is one of the most important components in multi-script document image analysis and it is used for a wide range of applications such as automatic storage of multi-script document images, document image retrieval, video indexing and retrieval, document sorting in digital libraries [40], [62]. Spitz [119] carried out the first extensive research on automatic script identification in 1994. Successively, two comprehensive surveys were conducted by Ghosh and Shivaprasad in 2000 [39] and Pal in 2006 [78]. Moreover, a survey specifically devoted to Indian script identification was published by Pal and Chaudhuri in 2004 [77]. It is worth noting that many researches on script identification have been devoted to Indian scripts since script identification is essential in a multi-lingual, multi-script country like India, where 18 official Indian languages and 12 different scripts are used. Documents are printed in three languages: English, Hindi (Devnagari) and the official regional language. More recently, several international competitions were also performed on script identification tasks [106], [109].

This paper presents a comprehensive survey of different script identification techniques. Section II gives a brief description of different writing systems. Section III introduces different document types and highlights the main discriminating methods. Section IV and V discuss the most profitable features and the classification techniques for script identification, respectively. System performances are analyzed in Section VI. Section VII presents some valuable trends for future research. Section VIII reports the conclusions of the paper.

II. SCRIPT WRITING SYSTEMS

In the world there are six large writing systems [39], [78]. Each writing system includes one or more scripts and each script can be used in one or more languages, as Table 1 shows.

1) *Logographic system.* A logographic system usually represents complete words. Han script is included in this system and used in Chinese, Japanese and Korean writings. This script can be distinguished from other Western and Asian scripts by its multiple short strokes, optical character density and appearance-based visual features.
2) *Syllabic system.* In this system, each symbol represents a syllable. Japanese scripts use a mix of logographic Kanji and syllabic Kanas and are part of this system. In these scripts, the presence of the simpler Kanas in between the logograms is less dense than in Chinese scripts and it is the distinguishing characteristic between Japanese and Han scripts.
3) *Alphabetic System.* The most important scripts in the alphabetic system are Greek, Latin, Cyrillic, and Armenian. The Greeks were the first Europeans to learn to write with an alphabet and from this system alphabetic writing spread to the rest of Europe, eventually leading to creation of all modern European alphabets. Latin script is used in many languages throughout the world such as Latin, Cyrillic, and Armenian, as well as many European languages like English, Italian, French, German, Portuguese, Spanish, and Austronesia, Modern Malay, Vietnamese, and Indonesian. Cyrillic script is used in some Eastern European, Asian, and Slavic languages such as Bulgarian, Russian, Macedonian, and Ukrainian. Some characters in the Cyrillic alphabet are borrowed from Latin and Greek and modified with cedillas, crosshatches, or diacritical marks.
4) *Abjads.* In this system, each symbol represents a consonantal sound. It includes Arabic and Hebrew scripts. The characteristic that clearly identifies Abjad-based scripts in pen computing systems is the right to left writing direction.
5) *Abugidas.* This system includes the Brahmic family of scripts which originated from the ancient Indian Brahmi script and makes up almost all of the scripts of India and Southeast Asia. The northern group of Brahmic scripts is used in the Devnagari, Bangla, Manipuri, Gurumukhi, Gujrati, and Oriya languages; the scripts in south India and southeast Asia are used in the Tamil, Telugu, Kannada, Malayalam Thai, Lao, Burmese, Javanese, Balinese languages. One important characteristic of Devnagari, Bangla, Gurumukhi, and Manipuri is that words are written together without spaces. The large number of horizontal lines in the textual portions of a document can distinguish these scripts from others.
6) *Featural system.* This system includes features that make up phonemes and includes the Korean Hangul

TABLE 1. Script writing systems.

Writing system	Scripts	Languages	Sample writings
Logographic system	Han	Chinese	基于局部特征进行识别的算法
		Japanese (Kanji)	とが多く、冬季の 影響するため、
		Korean (Hanja)	을 구성하여 상호 연 [illegible] 연결함으로
Syllabic system	Kanas	Japanese	とが多く、冬季の 影響するため、
Alphabetic system	Greek	Greek	[illegible]
	Latin	Afrikaans	Kan u dit weer sê?
		Catalan	Quina hora és?
		Dutch	Ik moet mijn Nederlands oefenen.
		English	Green is my favorite colour.
		French	Je vais vous emmener à l'hôpital.
		German	Bis spatter, Auf Wiedersehen! Tschuß!
		Italian	Attualmente gli strumenti di identificazione
		Luxembourgish	Wëllt schwätzt ëmwiesselen
		Portuguese	Falo um pouco de Português.
		Spanish	España Es Un País Maravelloso.
		Vietnamese	Viêm tiểu phế quản là m Trong hầu hết trường hợ
	Cyrillic	Russian	Мы пили напитки на балконе гостиницы
	American	English	Green is my favorite color.
Abjads	Hebrew	Hebrew	האדם נולדו בני חורין ושווים בערכם
	Arabic	Arabic	فلا يهلك كل من يؤمن به، بل تكون له الحياة الأبدية
		Farsi	واقعی ایران ایفا می کند و سه منطقه
		Urdu	[illegible]
		Uyghur	[illegible]
Abugidas	Brahmi scripts	Devanagari	[illegible]
		Bangla	দিয়েছে । কোম্পানীর পাঠ
		Manipuri	[illegible]
		Gurumukhi	[illegible]
		Gujrati	[illegible]
		Oriya	[illegible]
		Tamil	[illegible]
		Telugu	[illegible]
		Kannada	[illegible]
		Malayalam Thai	[illegible]
		Lao	[illegible]
		Burmese	[illegible]
		Javanese	[illegible]
		Balinese	[illegible]
Featural system	Hangul	Korean	을 구성하여 상호 연 [illegible] 연결함으로

script. In this system a script is formed by mixing logographic Hanja with featural Hangul. The Korean script is less complex and less dense compared to Chinese and Japanese scripts and it contains more circles and ellipses.

The six writing systems mentioned above include many scripts with similar shaped characters: similar shaped characters are the major source of confusion in script identification. Usually, each script has several spatial characters, diacritics, multi graphs (include digraphs) or ligatures that differ with other scripts in a same writing system. They are significant in identifying scripts with similar shaped characters each other. For instance, the 11 languages such as, Afrikaans, Catalan, Dutch, English, French, German, Indonesian, Luxembourgish, Malay, Portuguese, Spanish alphabets consisted of Latin alphabet based 26 characters, and they are distinguish together with spatial characters, diacritics, multi graphs and ligatures. In particular, the 26 Latin characters are included in all language alphabets and only three of them (English, Indonesian and Malay) are without diacritics. Other 8 types of language alphabets (Afrikaans, Catalan, Dutch, French, German, Luxembourgish, Portuguese, Spanish) contain several diacritics. These diacritics can be used to distinguish the 8 kinds of scripts in character level script identification. On the other hand, some diacritics are common for a number of scripts. For example, the diacritic ''é'' is not useful for character level script identification since it is common for Afrikaans, Catalan, Dutch, French, Luxembourgish, Portuguese and Spanish. In this case, other factors of these languages should be considered, such as multigraphs and ligatures.

A comparison of multigraphs and ligatures in the 11 Latin alphabet based scripts shows that Afrikaans, Catalan, Dutch, and Luxembourgish alphabets have no multigraphs and ligatures. English, French, German, Indonesian, Malay, Portuguese and Spanish alphabets contain several multigraphs. Ligatures are present in English, French and German alphabets. Therefore, these multigraphs and ligatures can be used as significant factors to identify these types of scripts in character level script identification. It is worth noting that some diacritics are not useful since they are common for several scripts. For example, the digraph ''ch'' is common for English, French, German, Portuguese and Spanish alphabets.

Arabic, Persian and Uyghur characters are similar to each other: they have 18 common characters. Moreover, Arabic and Persian alphabets have 8 common characters, Arabic and Uyghur have 2 common characters, Persian and Uyghur have 6 common characters. There are 6 Uyghur characters different from Arabic and Persian. Thus, character level script identification is not efficient to the 3 scripts and word/connected component level script identification should be considered.

III. SCRIPT IDENTIFICATION METHODS

Most research in the field of script identification concerns documents either printed or handwritten scripts. However, since several documents may contain text blocks with both printed and handwritten scripts, some research is now addressing hybrid documents. Hence, on the basis of content type, documents can be classified into three categories: printed, handwritten and hybrid.

Furthermore, document acquisition can be performed not only through optical scanners but also via cameras and camcorders. Of course, the device used for document acquisition can affect document image quality and therefore specific script identification methods have recently been proposed for video and camera-based acquisitions.

Table 2 summarizes some of the most relevant research in the field of script identification, categorized by document type and acquisition device. Earlier research on script identification of printed, handwritten and hybrid documents is discussed in the following. For each document type, different methods presented in the literature are introduced according to the kind of data they use to perform script identification: Page/Paragraph/Text-block, Text line, Word, or Character.

TABLE 2. Script identification: documents types vs acquisition.

	Scanner	Video and Camera
Printed	Chanda et al. [13, 14, 15, 16], Chaudhuri and Pal [21], Dhanya et al. [33], Ding et al. [34], Dhandra et al. [28, 29], Ghosh and Chaudhuri [41], Hochberg et al. [46], Jaeger et al. [48], Padma et al. [68], Pal et al. [71, 72, 73], Peake et al. [85], Prakash et al. [90], Rani et al. [91], Spitz [119]	Gllavata et al. [42], Linlin Li and Tan [54], Sharma et al. [104], Trung et al. [123], Zhou et al. [134]
HandPrinted	Chaudhuri [22], Dhandra and Hangarge [30], Ghosh et al. [39, 40], Hiremath et al. [44], Hochberg et al. [46], Namboodiri and Jain [64], Obaidullah et al. [67], Pal et al. [75], Razzak et al. [93], Roy et al. [96, 97, 98, 99], Zhou et al.[133]	
Hyb.	Ben Moussa et al. [6], Benjelil et al. [7], Kanoun et al. [50], Saidani et al. [101]	

A. SCRIPT IDENTIFICATION IN PRINTED DOCUMENTS

Most script identification research was carried out on printed documents. The main sources of printed documents include books, magazines, journals, dictionaries, etc.. Some researcher prepared multi script texts using automatic translating software [36] or other software [5] firstly, then multi script documents are obtained as computer printouts. Since the diversity of scripts and the deficiency of available public databases, most of the researcher built their own databases/datasets, as Table 3 summarizes.

1) PAGE/PARAGRAPH/TEXT BLOCK LEVEL SCRIPT IDENTIFICATION

Most research on printed document script identification has been carried out at the page level. Hochberg et al. [46] used cluster-based templates for discriminating 13 different scripts. Spitz [119] proposed a language identification

TABLE 3. Script identification: features.

	Level	Script/language	Data type and size	researcher
Printed document	Text line level	Bangla, Devanagari, English, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, Telugu, Urdu	3080 document images	S. Ghosh et.al [41] ; 2011
		Kannada, Hindi, Malayalam, Tamil, Telugu, Urdu and English	data set of 500 document images, which could contain 900 text lines	M.C. Padma et.al [68] ; 2009
		English , Chinese , Arabic , Devnagari and Bangla	700 different document images containing about 25000 text lines	U. Pal et.al [72] ; 2001
		Kannada , Gujrathi , Gurumukhi , Devnagari	250 different multi-script document images containing 4000 text lines	U. Pal et.al [73] ; 2003
	Word level	Telugu, Devnagari and English	1500 data (Telugu 600 , Devnagari 550 and English 350)	B.V. Dhandra et .al [28] ; 2006
		Kannada, Devnagari, English	1850 word images (Kannada 750, Devnagari 750 and English 350)	B.V. Dhandra et.al [29] ; 2006
		Arabic ,Chinese , Korean , Hindi	Not mentioned	S. Jaeger et.al [48] ; 2005
	Character level	English and Gurmukhi	19448 Characters and digits; 300dpi	Rani, R. et.al [91]; 2013
		Sinhala,Tamil,English	9450 data (consisting of 3913 Sinhala, 2582Tamil and 2955 English words); 300 dpi	Chanda et al. [15] ; 2008
HandPrinted document	Text block level	Bengali, English, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu	4000 images, 150 dpi	P.S. Hiremath et.al [44]; 2010
	Text line level	Bangla, Persian and Roman	Bangla 515 lines, Persian 537 lines and Roman 857 lines word, 150 dpi	Miguel A. Ferrer et.al,[37], 2014
	Word level	Bangla and English	4342 (3100 Bangla and 1242 English) handwritten words and 650 (400 Bangla 250 English) printed words	K. Roy et.al [96] ; 2005
		Bangla, Devnagari, Malayalam, Urdu, Oriya , Roman	160 script (32 Bangla, 30 Devnagari, 20 Malayalam, 16 Urdu, 30 Oriya, 32 Roman. 300 dpi.	K. Roy et.al [98] ; 2011
		Persian and Roman	5000 handwritten words (2577 images of Persian and the rest are Roman).	K. Roy et.al [99] ; 2010
		Bangia, Devanagari, Gurumukhi, Malayalam, Oriya ,Telugu ,Roman	dataset of 7000 handwritten text words	Pawan Kumar Singh et.al [115]; 2015
	Character level	Bangla, Devanagari, Roman and Urdu	Total 4000 numeral word image	Sk Md Obaidullah et.al [65]; 2015
Hybrid	Text line level	Arabic and Latin	2400 text images	S.Ben Moussa et. al [6]; 2008
	Word level	Arabic, Cambodian, Chinese, English, Greek, Hebrew, Japanese, Kannada, Korean, Mongolian, Russian, Thai ,Tibetan	16291 wild text images in 13 language scripts	Baoguang Shi et.al [111]; 2016
Video and Camera based data	Page out /Text block level	Arabia, Chinese, English, Hindi, Thailand and Korean	1800 images (Each script class has 300 gray images with 256*256 pixels) .	L. Zhou et.al [134] ; 2010
		Arabic, Chinese, English, Japanese, Korean and Tamil	1200 frames (Each language script has 200)	P. Shivakumara et.al [113]; 2015
	Text line level	English , chinese , Tamil	500 text images (English 200 , chinese 150 , Tamil 150 text lines)	Q.P. Trung et.al [123] ; 2011
	Word level	Hindi, English, Bengali	1271 words (430 Hindi, 410 English and 431 Bengali words)	Nabin Sharma et.al,[107]; 2014

scheme where the words of 26 different languages were first classified into Han-based and Latin-based scripts. Successively, the actual languages were identified using projection profiles of words and character shapes. Jie Ding et al. [34] presented a method which uses a combined analysis of several discriminating statistical features to categorize European and Oriental language scripts. Chaudhuri and Pal [21] developed a system for identifying Bangla and Devnagari (Hindi) scripts using a classification tree. Research on printed document script identification was also conducted at text block level. For instance, Peake and Tan [85] proposed a method based on Multiple Gabor filters and grey level co-occurrence matrices to extract the texture features of five major scripts.

2) TEXT-LINE LEVEL SCRIPT IDENTIFICATION

In text-line level script identification, a text block is firstly divided into lines. Pal and Chaudhuri [71] developed an automatic technique for separating text lines using script characteristics and shape based features. They also proposed a system for the identification of printed Roman, Chinese, Arabic, Devnaguri and Bangla text lines from a single document [72] and a method of identifying text lines of different Indian scripts from a document [73]. An automatic technique for the identification of Japanese and English script portions from a single line of a printed document was proposed by Chanda et al. [13]. Padma and Vijaya [68] developed a monothetic algorithmic model to identify and separate Telugu, Hindi and English text lines from a printed multilingual document. A simple and efficient technique of script identification for Kannada, Hindi and English text lines was presented by Prakash et al. [90]. Ferrer et al. [36] proposed a LBP-based line-wise script identification system to identify ten different scripts.

3) WORD LEVEL SCRIPT IDENTIFICATION

Dhanya and Ramakrishnan [33] presented a successful method for identifying script at word level in a bilingual document containing Roman and Tamil scripts. Jaeger et al. [48] used a Gabor filter analysis of textures and a multiple classifier system with four different classifiers to identify Arabic, Chinese, Hindi, and Korean scripts at word-level. Dhandra et al. [28], [29] proposed an automatic technique for script identification at word level based on the morphological reconstruction of two printed scripts: Telugu and Devnagari. A SVM based method was proposed by Chanda et al. [14] for the identification of printed English and Thai scripts at word level from a single line of a document page. Chanda *et al.* [15] proposed a SVM based technique for word-level identification of Sinhala, Tamil and English scripts from a single document page, and a SVM based scheme for the identification of printed word-level English, Devnagari and Bangla scripts [16].

4) CHARACTER LEVEL SCRIPT IDENTIFICATION

Pal and Sarkar [74] used a combination of topological, contour and water reservoir concept based features to identify printed Urdu script. Rani et al. [91] carried out experiments on multi-font and multi-sized characters with Gabor features and Gradient features to identify Gurumukhi and English scripts at character or numeral level.

B. SCRIPT IDENTIFICATION IN HANDWRITTEN DOCUMENTS

Handwritten documents are another important area of application for script identification systems. Of course, script identification of handwritten documents is more challenging than script identification of printed documents. In fact, there are some relevant differences between printed and handwritten script identification. For example, some scripts resemble each other much more in handwritten documents than in printed ones. Moreover, handwriting styles can be very variable. The experimental documents, which are written by different individuals at different times, enlarge the inventory of possible character and word shapes in handwritten documents. In addition, ruling lines and character fragmentation are common in handwritten documents due to the variety of papers and writing instruments used. All these differences can create huge challenges for script identification in handwritten documents.

1) PAGE/PARAGRAPH/TEXT BLOCK LEVEL SCRIPT IDENTIFICATION

The first study conducted on handwritten script identification was carried out by Chaudhuri [22] and was similar to that proposed by Hochberg et al. [46] for printed documents. However, the resulting classification accuracy was lower than that for the printed documents. An online handwritten script recognition system was proposed by Namboodiri and Jain [64] for classifying six major scripts at word level. Eleven different features and six types of classifiers were considered. A method based on the texture features for script identification in a handwritten document image was proposed by Hiremath et al. [44]. Ghosh and Shivaprasad [39] proposed an handwritten script identification method in which a ''possibilistic'' approach was used for cluster analysis.

2) TEXT-LINE LEVEL SCRIPT IDENTIFICATION

Namboodiri and Jain [64] proposed a method to classify words and lines into one of the six major scripts: Arabic, Cyrillic, Devnagari, Han, Hebrew or Roman. The classification is based on eleven different spatial and temporal features extracted from strokes of the words.

3) WORD LEVEL SCRIPT IDENTIFICATION

Roy et al. [96] proposed a word-wise handwritten script identification method for Indian postal automation regarding Bangla and English script identification at word-level. The method mainly uses water reservoir concept based features, fractal-based features and a Neural Network classifier. Roy and Majumder [97] also developed a technique for script separation of handwritten postal documents in Bangla, Roman and Devanagri scripts. Run Length Smoothing Algorithm (RLSA) was used to segment the document pages into lines and then into words. Fractal-based, busy-zone and topological features were used along with a Neural Network (NN) classifier for script identification. A script separation technique of Roman and Oriya scripts for Indian Postal automation was proposed by Zhou et al. [133]. They presented a script identification method based on water reservoir concept based features, fractal dimension based features and topological features with an NN classifier. Sarkar et al. [103] presented an automatic separation system for word-level script identification from Bangla or Devanagri mixed with Roman scripts. Dhandra and Hangarge [30] used a two-stage approach. In the first stage, some global and local features were applied to identify the text words. In the second stage, the numeral written in different scripts was identified. To test the system, Kanada, Devanagri and Roman scripted handwritten documents were considered. A word-wise handwritten script identification system for bi-script documents written in Persian and Roman scripts was proposed by Roy et al. [99]. The system used a simple and fast computable sets of twelve features based on fractal dimension, position of small components and topology. A scheme for document level handwritten script identification from six popular Indian script documents was presented by Roy et al. [98]. In the proposed scheme, a small set of features based also on fractal dimension are computed using an MLP classifier. Obaidullah et al. [67] proposed a scheme to identify the six popular Bangla, Devnagari, Malayalam, Urdu, Oriya and Roman scripts in Indian documents, and compared performance using different well-known classifiers.

4) CHARACTER LEVEL SCRIPT IDENTIFICATION

Pal et al. [75] proposed a modified quadratic classifier based scheme for the recognition of off-line handwritten numerals of six popular Indian scripts: Devnagari, Bangla, Telugu, Oriya, Kannada and Tamil. Razzak et al. [93] presented a fuzzy rule based approach for the recognition of both Urdu and Arabic numerals in an unconstrained environment.

C. SCRIPT IDENTIFICATION IN HYBRID DOCUMENTS

Hybrid documents include printed and handwritten texts. A multi-lingual automatic identification of Arabic and Latin in both handwritten and printed script was proposed by Ben Moussa et al [6]. A method for Arabic and Latin text block differentiation for both printed and handwritten scripts is discussed by Kanoun et al. [50]. The method is based on a morphological analysis for each script at the text-block level and a geometrical analysis at line and connected component levels. Benjelil et al. [8] proposed an accurate system based on a steerable pyramid transform for Arabic and Latin script identification at word level. By using new structural features, a successful attempt was made by Saidani et al. [101] to identify the Arabic or Latin script of a machine printed or handwritten document at word level.

D. SCRIPT IDENTIFICATION IN VIDEO FRAMES AND CAMERA BASED IMAGES

The extraction of script information from video frames or camera based images has not been much explored so far. Unlike printed or handwritten documents, video and camera based script identification methods first require the extraction of textual information: this is an important and very complex task. In printed and handwritten documents, text in black appears generally on a simple background (colorless). However, script recognition in video and camera based images originates from complex conditions and suffer from low resolution, blur, complex background, noise, orientation problems, different fonts and font sizes of video text, etc. All these complications make this problem more difficult and challenging than printed and handwriting document identification. Some approaches on script information from video frames or camera based images at different levels are reported in the following.

1) PAGE/PARAGRAPH/TEXT BLOCK LEVEL SCRIPT IDENTIFICATION

Gllavata and Freisleben [42] presented an approach for discriminating between Latin and Ideographic scripts by a set of low-level features. The decision is made using a K-Nearest Neighbour classifier. New Spatial-Gradient based Features (SGF) were proposed by Zhao et al. [132] for script identification at block level for six scripts namely, Arabic, Chinese, English, Japanese, Korean and Tamil.

2) TEXT-LINE LEVEL SCRIPT IDENTIFICATION

Phanet al. [87] proposed two features, namely smoothness and cursiveness, for video script identification at text-line level. In their approach, English, Chinese and Tamil scripts were considered.

3) WORD LEVEL SCRIPT IDENTIFICATION

Sharma et al. [104] used Zernike moments, Gabor and gradient features with SVM classifiers to identify English, Bengali and Hindi scripts. A study of word level multi script identification from video frames is proposed by Sharma et. al. [107] using different combinations of texture based features namely, Local Binary Pattern (LBP), Gradient, Histogram of Oriented Gradient (HoG) and Gradient Local Auto-Correlation (GLAC) features. SVMs and ANNs classifiers were applied for English, Bengali and Hindi scripts identification. This experiment pointed out the efficiency of gradient features for low resolution, blur, complex background, and noise video based images. Shivakumara et. al [112] developed a word level script identification method for Arabic, Chinese, English, Japanese, Korean and Tamil scripts by using new Gradient Angular Features. Bag-of-Visual Words based word-wise script identification from video images is presented by Sharma et al. [108] for five different south Indian scripts.

4) CHARACTER LEVEL SCRIPT IDENTIFICATION

Li and Tan [54] reported a script identification based on statistical features technique to identify character level English, Arabic and Chinese scripts of camera-based images. The experimental results show that this method is tolerant to moderate perspective variations and document skew.

IV. FEATURES FOR SCRIPT IDENTIFICATION

Feature extraction is a vital part of any practical recognition system. In the last few years, different kinds of features have been evaluated for script identification based on the characters of each script.

Two broad categories of features have been established in the script identification field, as Table 4 summarizes [26]: local feature and global feature. Local features are extracted from small textual components of the document image. Therefore, they strongly depend on the effectiveness of the segmentation procedure. Statistical-, structural- and template-based characteristics are examples of local features [26]. Global features are extracted from blocks of text of the document image. Texture- and Steerable pyramid-based features are examples of global features [26].

A. LOCAL FEATURES

The analysis of local features mainly considers the analysis of intrinsic features such as character shape based features, structural features, statistical features, morphological, topological and contour based features, water reservoir principle

TABLE 4. Script identification: features.

Features			Granularity of data	References
Global features		Steerable pyramid feature	Word level	Benjelil et al. [7, 8]
Global features	Texture features	Gabor filters	Page out/text block level	Busch et al. [11, 12], Chaudhury and Sheth [23], Nagabhushan et al. [63], Pan et al. [80], Pati et al. [82], Peake and Tan [85], Singhal et al. [116], Tan [121]
			Word level	Dhanya and Ramakrishnan [33], Jaeger et al. [48], Pati et al. [83], Sanjeev and Sudhaker [102]
			Character level	Chanda et al. [18], Philip and Samuel [88], Rani et al. [91]
		Grey level co-occurrence matrix	Page out/text block level	Busch et al. [11], Pan et al. [80], Peake and Tan [85]
			Text line level	Ablavsky and Stevens [1]
			Word level	Ablavsky and Stevens [1]
				Sharma et al. [104]
		Wavelet feature	Page out/text block level	Busch et al. [11], Hiremath et al. [44, 45], Padma and Vijaya [69], Zhou et al. [134]
			Word level	Angadi and Kodabagi [3]
			Text line level	Ding et al. [34], Rezaee et al. [94], Waked et al. [126]
			Character level	Chanda et al. [14, 17]
		Connected component feature	Page out/text block level	Chaudhury and Sheth [23]
			Text line level	Pal et al. [76]
			Word level	Hochberg et al. [46], Obaidullah et al. [67], Roy et al. [95, 98]
			Character level	Peng et al. [86]
		Project profile	Page out/text block level	Bashir and Quadri [5], Lam et al. [52]
			Text line level	Ben Moussa et al. [6], Chaudhury and Sheth [23], Ding et al. [34], Pal and Chaudhuri [71], Pal et al. [72, 73], Pal et al. [76], Prakash et al. [90], Waked et al. [126]
		Project profile	Word level	Dalal and Malik [26], Dhanya and Ramakrihnan [33], Rezaee et al. [94], Roy et al. [96], Saidani et al. [101]
			Character level	Chanda et al. [13]
		Upward concavities	Text line level	Khoddami et al. [51]
			Word level	Dalal and Malik [26], Rezaee et al. [94], Roy et al. [96], Spitz [119]
		Water reservoir based feature	Text line level	Ghosh and Chaudhuri [41], Pal et al. [72, 73], Pal et al. [76]
			Word level	Roy et al. [95, 96]
			Character level	Chanda et al. [14, 15, 17], Pal et al. [74]
		Fractal based feature	Page out/text block level	Tho and Tang [122]
			Text line level	Ben Moussa et al. [6]
			Word level	Lam et al. [52], Obaidullah et al. [67], Roy et al. [96, 97, 98]

TABLE 4. *Continued.* Script identification: features.

	Structural/Geometrical features	Head-line (heuristics) feature	Text line level	Pal and Chaudhuri [71],
			Word level	Chanda et al. [17], Das et al. [27], Roy et al. [95, 96, 98], Saidani et al. [101]
		Moment (boundary) features	Text line level	Ablavsky and Stevens [1], Gllavata and Freisleben [42]
			Word level	Ablavsky and Stevens [1], Chanda et al. [18], Sharma et al. [104]
		Morphological (stroke) feature	Page out/text block level	Kanoun et al. [50]
			Text line level	Not used yet
			Word level	Chanda and Pal [17], Dhandra et al. [31], Namboodiri and Jain [64],
			Character level	Dhandra et al. [32], Shijian Lu and Tan [56] Peng et al. [86]
		Topological (loop) feature	Text line level	Pal et al. [73]
			Word level	Chanda et al. [15], Dalal and Malik [26], Lam et al. [52], Roy et al. [96, 97, 98], Saidani et al. [101]
	Template Matching feature		Page out/text block level	Shijian Lu and Tan [94], Wang et al. [127]
			Word level	Hochberg et al. [46]
			Character level	Chaudhuri and Pal [21], Ghosh and Shivaprasad [39]

based features, etc. The extraction of these features is time consuming, but they convey relevant characteristics for script identification [26], [30], [69], [112]:

1) STATISTICAL FEATURES

Statistics-based features extract mathematical characteristics as the mean and variance of the width, height, ratio and area of the connected components. They concern methods that identify scripts through the analysis of the upward concavity, vertical and horizontal projections, etc. These methods are more suitable to scripts that differ significantly in style. Statistics-based approaches are highly sensitive to noise and image quality and all features are extracted at higher levels such as words, lines and text blocks. Some of the commonly used statistical features are [52]: Horizontal projection profiles [8], [17], [23], [24], [27], [33], [34], [52], [72]–[74], [76], [77], [90], [101]; Water reservoir-based features [6], [13], [17], [74], [76], [95], [96]; Bounding box feature [13], [94], [118], [126]; Character pitch features [13]; Upward concavities [53], [118], [119].

2) STRUCTURE/GEOMETRIC FEATURES

Structural features include loops, cusps, endpoints, starts points, etc.. Structural features depend on the instinctive aspects of writing and are based on the geometric appearance of scripts. Some typical structural features are [11], [17], [73], [96]: Head-line (heuristics) features [17], [27], [71], [73], [96], [99]; Fractal-based features [6], [67], [122]; Topological features [15], [97], [99]; Morphological features [17], [28], [29], [31], [56], [64].

3) TEMPLATE MATCHING FEATURES

In the template-based approach, an unknown pattern is superposed directly on the ideal template pattern and the degree of correlation between the two is used for classification. In general these methods have advantages in distinguishing similar scripts although they are strongly sensitive to the font and size variations of characters [21], [38], [46], [127].

B. GLOBAL FEATURES

Global features for script identification are based on DCT, DWT, Gabor, steerable pyramids, and Radon transform [19], [22], [43], [70], [75], [79], [89], [97], [120]. These are robust to noise, small skew, and faster in computation than local features. In general, global features are considered to be efficient in characterizing large size texture patterns, e.g. text blocks. Furthermore, since these features regard a text block as one single entity, analysis at the levels of text lines, words or connected components is not possible [26, 30]. Typical global features are derived from texture analysis approaches and include Gabor Filter [59], [66], [82], [85], [87], [88], [92], Wavelet Transform features [3], [44], [73], Discrete Cosine Transform (DCT) [92], Gray level co-occurrence matrix [9], [44], rotation invariant features [85], gradient features [16], [91], [104], [132], steerable pyramid transforms [7], [8], etc..

V. CLASSIFICATION TECHNIQUES FOR SCRIPT IDENTIFICATION

Although classification is a crucial step of script identification systems, the literature shows that only a few simple classifiers were used in earlier works, as Table 5 reports [21], [31], [34], [46], [118], [119], [129]. The K-Nearest Neighbor (K-NN) classifier has been extensively used in script identification systems based on Gabor filter [81], [85], Cartesian moments [1], appearance based model approaches [125], grey level co-occurrence matrix features [85], statistical-based features [20], stroke density and distribution-based features [56], texture features [20], [42]. Support vector machine (SVM) has also been applied to script identification. SVM-based systems for script identification use structural features, topological features and water reservoir principle based features [14], [15], Zernike moment–based feature [18], [104], Gabor, and gradient features [104].

Other classification methods were considered for script identification such as Neural Network [9], [96], quadratic classifier [43], [64], [75], [126], [134], rule-based classifiers [1], [90], [92], [93], Linear Discriminant Classifiers [43], [55], [83], Gaussian Mixture Model [11], [48], [50], [99], etc..

VI. PERFORMANCE ANALYSIS

Performance of some of the most significant systems presented in the recent literature are reported in Table 6. More precisely, Table 6 shows that different kinds of features, e.g., statistical features, structural features, symbols matching features and texture features, were generally used for script identification. Indeed, these features express only some characteristics of scripts that are generally not sufficient for script identification. Conversely, texture features are generally more efficient than others, but they cannot be applied reliably at word and character level within a document. Furthermore, to achieve better results, different kinds of features and classifiers were used. Experimental results show that a particular feature, that is generally efficient within a set of scripts, is not necessarily useful for other scripts. In English, Kannada, Hindi multi-script documents [32], [69], [81], texture features with a K-NN classifier were more useful than other features. For identifying different scripts in Indian documents, whatever classifier was used, both Global and Local features demonstrated to be efficient. Moreover, results obtained using a single feature (either local or global) were generally worse than those obtained using both features [31],

TABLE 5. Script identification: classifiers.

Classifiers		References
Binary Tree classifier	Binary tree	Chanda et al. [13, 17], Roy et al. [101]
	Decision tree	Lin et al. [55], Namboodiri and Jain [64], Rani et al. [91]
Gaussian Mixture Model (GMM)		Busch et al. [11], Jaeger et al. [48]
Linear Discriminant classifier (LD)		Hangarge et al. [43], Pati an Ramakrishnan[83]
Nearest Neighbor (KNN)	K-Nearest Neighbor (K-NN)	Ablavsky et al [1], Ben Moussa [6], Benjelil et al. [7], Chaudhari and Gulati [20], Dhanya and Ramakrishnan [33], Dhandra et al. [28, 29, 30], Ghosh and Chaudhuri [41], Gllavata and Freisleben [42], Hangarge et al. [43], Hirenath et al. [44], Jaeger et al. [48], Kanoun et al. [50], Khoddami et al. [51], Namboodiri and Jain [64], Padma and Vijaya [69], Peake and Tan [85], Rani et al. [92], Roy et al. [96], Shijian Lu and Tan [57], Trung et al. [123], Zhao [132]
	Minimum distance	Ghosh and Chaudhuri [41]
	Euclidean distance (ED)	Dhanya and Ramakrishnan [33], Pati and Ramakrishnan [83], Jaeger et al. [48]
Neural Network (NN)	Multi-Layer Perceptron (MLP)	Bhattacharya and Chaudhuri [9], Namboodiri and Jain [64], Pan et al. [80], Roy et al. [96, 97, 98], Sarkar et al. [103]
	Bayesian	Namboodiri and Jain [64]
	Modular	Patil and Subbareddy [84]
	Probabilistic	Patil and Subbareddy [84], Rani et al. [92]
Quadratic Classifier	Modified Quadratic Discriminant Function	Pal et al. [75]
	(MQDF)	
	Bayes Quadratic classifier	Namboodiri and Jain [64]
Rule-based classifier		Prakash et al. [90], Razzak et al. [93]
Support Vector Machine (SVM)	Linear	Chanda et al.[14], Dhanya and Ramakrishnan [33], Ghosh and Chaudhuri [41], Sharma et al. [104], Wang et al. [127]
	Polynomial/Gaussian radial basis functions (RBFs)	Ben Moussa [6], Chanada et al. [14, 15, 16, 18, 19], Ferrer. et al. [36], Ghosh and Chaudhuri [41], Jaeger et al. [48], Namboodiri and Jain [64], Pan and Tang [79], Rani et al. [92], Sharma et al. [104], Wang et al. [127], Zhou et al. [134]

TABLE 6. Comparison results.

References	Script	Methods		Accuracy (%)
		Feature	Classifier	
Angadi and Kodabagi [3]	Hindi, Kannada, English, Malayalam, Tamil	Wavelet feature	NA	94.33
Bashir and Quadri [5]	English, Kashmiri	Project profile feature	NA	96.2
Ben Moussa et al. [6]	Arabic, Latin	Fractal feature	KNN, RBF	96.64 98.72
Benjelil et al. [7]	Arabic, Latin	Steerable pyramid transform	KNN	97.5
Chanda et al. [13]	Japanese, English	Statistical and structural feature	Tree Classifier	98.79
Chanda et al. [14]	English, Tail	Statistical and structural feature	SVM	99.36
Chanda et al. [15]	Sinhala, Tamil and English	Statistical and structural feature	SVM	96.4
Chanda et al. [16]	English, Devnagari, Bangla	Chain code histogram feature	SVM	98.51
Chanda et al. [17]	English, Devnagari, Urdu	Statistical and structural feature	Tree classifier	97.51
Chanda et al. [18]	Chinese, Japanese, Korean, Roman	Chain code histogram feature	SVM	98.39 (character) 99.85 (word)
Dhandra et al. [28]	English, Kannada, Hindi	Morphological feature	KNN	97
Dhandra et al. [30]	Kannada, Roman, Devnagari	Global and Local feature	KNN	99.96
Dhanya and Ramakrushnan [33]	Roman and Tamil	Gabor filter	SVM NN KNN	96.03 91.86 90.02
Kanoun et al. [50]	Arabic, Latin	Morphological and geometrical feature	KNN	96.1
Lin et al. [55]	Chinese, English	NA	SVM	99.6
Namboodiri and Jain [64]	Arabic, Cyrillic, Devnagari, Han, Hebrew, Roman	Stroke feature	KNN, NN, SVM	95.5
Padma and Vijaya [69]	English, Kannada, Hindi	Texture feature	KNN	99.33
Pal and Chaudhuri [71]	English, Devnagari etc.	Shape based feature	NA	98.5
Pal et al. [72]	Roman, Chinese, Arabic, Devnagari, Bengla	Statistical feature	NA	97.33
Pal et al. [74]	Indian	Statistical feature	NA	97.52
Pan et al. [80]	Chinese, Japanese, Korean, English	Gabor filter	NN	98.5
Patil and Subbareddy [84]	English, Kannada, Hindi	NA	KNN PNN	98.0 98.89
Pati and Ramakrishnan [83]	Indian	Gabor filter and DTC	NN, SVM	99.6
Peake and Tan [85]	Chinese, English, Greek, Korean, Malayalam, Persian, Russian	Gabor filter, Grey level co-occurrence matrix	KNN	95
Philip and Samuel [88]	English, Malayalam	Gabor filter	NA	96.5
Rani et al. [91]	English, Gurumukhi	Gabor filter, Gradient feature	SVM	98.9 99.45
Rani et al. [92]	English, Gurumukhi	Structure, Gabor, DCT feature	SVM, KNN, PNN	99.402
Rezaee et al. [94]	Farsi, English	Statistical and shape based feature	NA	96.05
Roy et al. [96]	Indian	Statistical feature	NN	97.62
Roy et al. [97]	Indian	Statistical and structural feature	NN	96.79
Saidani et al. [101]	Arabic, Latin	Structure feature	NA	98.4
Shijian Lu and Tan [56]	Arabic, Chinese, Hebrew, Roman	Statistical feature	KNN	95.36
Singhal et al. [116]	Roman, Devanagari, Bangla, Telugu	Gabor filter	NA	91.6
Waked et al. [126]	Roman, Ideographic, Arabic	Statistical feature	NA	91
Wang et al. [127]	Chinese, Japanese, Korean, English	Template based method	SVM	99.1
Zhou et al. [134]	Arabic, Chinese, English, Japanese, Korean, Thailand	Texture feature	SVM, RBF	97.59

[83]. In Han, Roman and some alphabet multi script identification methods, an SVM classifier with different extracted features achieved higher identification results than K-NN and NN classifiers [19], [55], [127]. For Arabic and Latin scripts, an RBF classifier was more efficient than a K-NN classifier [6].

It is worth noting that approaches in Table6 are difficult to be compared since performance have been estimated using databases collected in laboratory environments. In fact, although there existed several public datasets containing natural images with texts, they are mainly related to the text recognition task [109], [128]. Some datasets specifically devoted to script identification have been also realized [46], [87], [132]. Among the others, the SIW-10 dataset was

TABLE 7. Public databases.

DB	Script Lang.	Data type and size	Special field	Reference	Year
IAM	English	Handwritten, 1539 page of text images written by 657 different writers	writer identification	U. Marti et. al [61]	2002
APTI	Arabic	Printed, 113284 words, 10 fonts, 10 sizes, 4 styles	evaluation of screen-based OCR systems	F. Slimane et. al [117]	2009
APTID/MF	Arabic	Printed,1845 text-blocks images, 27402 characters images	text recog., OCR systems	F.K. Jaiem et. al [49]	2013
KAFD	Arabic	Printed,40 Arabic fonts, 115068 page images, 2576024 line images	multi-font text recog.	H. Luqman et. al [58]	2014
RIMES	French	Handwritten mails, 1300 individuals, 12723 pages of 5605 letters	writer identification, handwriting recog.	E. Augustin et. al [4]	2006
KHATT	Arabic	Handwritten, 1000 writers 4000 image pages	text recog., writer identification, forms analysis, segmentation	S.A Mahmoud et. al [60]	2012
QUWI	English, Arabic	Handwritten, 1017 writers image pages	writer identification and gender, age and handedness classification	S.Al-Maadeed et. al [2]	2012
LAMIS-MSHD	French, Arabic	Handwritten 600 Arabic, 600 French handwritten, 1300 signatures, 21000 digits	Writer recog. writer classification, signature verification	C. Djeddi et. al [35]	2014
Maurdor	English, Arabic, French	Printed and handwritten, 2500 document images	script identification	S. Brunessaux et. al [10]	2014
SIW-10	Arabic, Chinese, English, Greek, Hebrew, Japanese, Korean, Russian, Thai and Tibetan	13045 word images, cropped from 7700, full images from 10 languages	script identification	Baoguang Shi et. al [110]	2015
APTID/MF	English, Arabic,	1. Printed Text images, English: 2328 Arabic: 1845; 2. Handwritten Text images, English: 582, Arabic: 460	script identification	I. Chtourou, A. Cheikh et. al [49]	2013

developed for script identification [110]. The SIW-10 dataset contains more than 13,000 multi-scripts images including textual components from 10 languages. Therefore, to date, it can be considered as one of the most valuable benchmarking dataset for research in script recognition.

VII. FUTURE WORK AND TRENDS

Although in the last twenty years there have been many advances in the field of script identification, a great deal of work is still necessary to improve accuracy and efficiency of script identification systems. Some of the most valuable directions of research are here addressed.

The first point is that, as Table 3 shows, many researchers constructed database/datasets by themselves collecting data in laboratory according to the requirement of their research work. These datasets are different in type, size, scanning resolution and image format. Besides, these datasets are not publicly available. Therefore, specific work is necessary to define some standard data formats and to realize and release public datasets for script identification.

Table 7 summarizes some of the public datasets that have been considered by the script identification research community. Most of these databases are referred to research in the field of OCR [117], [49], handwriting recognition [4], [58], [60], document analyzing [60], writer identification and classification [2], [4], [61], [60], signature recognition and verification [35]. Only few public databases are specifically devoted to script identification [110], [10], [49]. Moreover, databases for script identification are limited in terms of script/language type, font types and sizes, lack degraded/noisy images, etc.. In particular, many databases are devoted to specific scripts, such as English, Chinese, Arabic, German, French, Japanese, Korean, Devanagari, Bangla, whereas no databases are available for research on other scripts. Of course, along with the expansion of research on script identification, the blank of some scripts in this area will be amended in the future work.

Furthermore, most of researches is based on offline script identification technology, but there are only few reports about online script identification technology. With the spreading of PDAs and smartphones, the demand of online script identification technology is increasing. At the same time, for everyday useful aims, also automatic translation should be considered. Just think to a tourist needing for a fast automatic translation of a signboard. Of course the development of online script identification systems will require a great amount of work and it is an extraordinary challenge for the research community.

VIII. CONCLUSION

Script identification is an important task in an OCR system for multi-lingual, multi-script documents. Many script identification methods have been proposed for written scripts at different levels within a document— page/paragraph level, text-line level, word level, and even character level.

Compared to the field of document analysis and optical character recognition, research on script identification is still limited. In fact, studies were focused so far on identifying the major scripts in the world such as English, Arabic, Indian, Chinese and Japanese. Indeed, many other scripts exist that have received no apparent attention. The identification of video and camera based images is another research area for which more research is necessary, since mobile and cheap devices become more and more widespread.

Concerning features and classification methods, although it is quite difficult to obtain conclusive results, Gabor filter and statistical features are certainly some of the most effective characteristics for script identification as well as kNN and SVM are the most valuable classifiers.

It can be concluded that although many advancements have been made, additional research is necessary the field of script identification. A crucial step is certainly the creation of new standard databases for multi-lingual, multi-script identification, to evaluate comparatively different script recognition systems and support the research community active in the field.

REFERENCES

[1] V. Ablavsky and M. R. Stevens, "Automatic feature selection with applications to script identification of degraded documents," in *Proc. ICDAR*, Aug. 2003, pp. 750–754.

[2] S. Al Maadeed, W. Ayouby, A. Hassaïne, and J. M. Aljaam, "QUWI: An arabic and English handwriting dataset for offline writer identification," in *Proc. 13th ICFHR*, Sep. 2012, pp. 742–747.

[3] S. A. Angadi and M. M. Kodabagi, "A fuzzy approach for word level script identification of text in low resolution display board images using wavelet features," in *Proc. ICACCI*, Aug. 2013, pp. 1804–1811.

[4] E. Grosicki, M. Carré, J. M. Brodin, and E. Geoffrois, "Results of the RIMES evaluation campaign for handwritten mail processing," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, Barcelona, Spain, 2009, pp. 941–945.

[5] R. Bashir and S. Quadri, "Identification of Kashmiri script in a bilingual document image," in *Proc. ICIIP*, Shimla, India, Dec. 2013, pp. 575–579.

[6] S. Ben Moussa, A. Zahour, A. Benabdelhafid, and A. M. Alimi, "Fractal-based system for Arabic/Latin, printed/handwritten script identification," in *Proc. ICPR*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[7] M. Benjelil, S. Kanoun, R. Mullot, and A. M. Alimi, "Arabic and latin script identification in printed and handwritten types based on steerable pyramid features," in *Proc. ICDAR*, Barcellona, Spain, Jul. 2009, pp. 591–595.

[8] M. Benjelil, R. Mullot, and A. Alimi, "Language and script identification based on steerable pyramid features," in *Proc. ICFHR*, Bari, Italy, Sep. 2012, pp. 716–721.

[9] U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 444–457, Mar. 2009.

[10] S. Brunessaux *et al.*, "The maurdor project: Improving automatic processing of digital documents," in *Proc. DAS*, Apr. 2014, pp. 349–354.

[11] A. Busch, W. W. Boles, and S. Sridharan, "Texture for script identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1720–1732, Nov. 2005.

[12] A. Busch, "Multi-font script identification using texture-based features," in *Image Analysis and Recognition* (Lecture Notes in Computer Science), vol. 4142. A. Campilho and M. Kamel, Eds. Berlin, Germany: Springer, ICAR 2006, pp. 844–852.

[13] S. Chanda, U. Pal, and F. Kimura, "Identification of Japanese and English script from a single document page," in *Proc. IEEE-CIT*, Oct. 2007, pp. 656–661.

[14] S. Chanda, O. R. Terrades, and U. Pal, "SVM based scheme for Thai and English script identification," in *Proc. ICDAR*, Paraná, Argentina, Sep. 2007, pp. 551–555.

[15] S. Chanda, S. Pal, and U. Pal, "Word-wise Sinhala Tamil and English script identification using Gaussian kernel SVM," in *Proc. ICPR*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[16] S. Chanda, S. Pal, K. Franke, and U. Pal, "Two-stage approach for word-wise script identification," in *Proc. ICDAR*, Jul. 2009, pp. 926–930.

[17] S. Chanda and U. Pal, "English, devnagari and urdu text identification," in *Proc. Int. Conf. Cognit. Recognit.*, 2005, pp. 538–546.

[18] S. Chanda, fK. Franke, and U. Pal, "Identification of Indic scripts on torn-documents," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 713–717.

[19] S. Chanda, U. Pal, K. Franke, and F. Kimura, "Script identification—A han and roman script perspective," in *Proc. ICPR*, Istanbul, Turkey, Aug. 2010, pp. 2708–2711.

[20] S. A. Chaudhari and R. M. Gulati, "An OCR for separation and identification of mixed English—Gujarati digits using kNN classifier," in *Proc. ISSP*, Gujarat, India, Mar. 2013, pp. 190–193.

[21] B. B. Chaudhuri and U. Pal, "An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi)," in *Proc. ICDAR*, Ulm, Germnay, 1997, pp. 1011–1015.

[22] B. B. Chaudhuri, "On multi-script OCR system evaluation," in *Proc. Int. Workshop Perform. Eval. Issues Multi-Lingual (OCR)*, 1999, p. 1. [Online]. Available: http://www.kanungo.com/workshop/abstracts/chaudhuri.html

[23] S. Chaudhury and R. Sheth, "Trainable script identification strategies for Indian languages," in *Proc. ICDAR*, Sep. 1999, pp. 657–660.

[24] J. Cheng, X. Ping, G. Zhou, and Y. Yang, "Script identification of document image analysis," in *Proc. 1st Int. Conf. Innov. Comput., Inf. Control*, Beijing, China, Aug. 2006, pp. 178–181.

[25] I. Chtourou, A. C. Rouhou, F. K. Jaiem, and S. Kanoun, "ALTID: Arabic/Latin text images database for recognition research," in *Proc.13th ICDAR*, Aug. 2015, pp. 836–840.

[26] S. Dalal and L. Malik, "A survey of methods and strategies for feature extraction in handwritten script identification," in *Proc. Int. Conf. Emerg. Trends Eng. Technol. (ICETET)*, Nagpur, India, Jul. 2008, pp. 1164–1169.

[27] M. S. Das, D. S. Rani, and C. R. K. Reddy, "Heuristic based script identification from multilingual text documents," in *Proc. Int. Conf. Recent Adv. Inf. Technol. (RAIT)*, Dhanbad, India, Mar. 2012, pp. 487–492.

[28] B. V. Dhandra, H. Mallikarjun, R. Hegadi, and V. S. Malemath, "Word-wise script identification based on morphological reconstruction in printed bilingual documents," in *Proc. IET Int. Vis. Inf. Eng. (VIE)*, Bangalore, India, Sep. 2006, pp. 389–393.

[29] B. V. Dhandra, H. Mallikarjun, R. Hegadi, and V. S. Malemath, "Word-wise script identification from bilingual documents based on morphological reconstruction," in *Proc. Int. Conf. Dig. Inf. Manage.*, Bangalore, India, Dec. 2006, pp. 389–394.

[30] B. V. Dhandra and M. Hangarge, "Global and local features based handwritten text words and numerals script identification," in *Proc. ICCIMA*, Dec. 2007, pp. 471–475.

[31] B. V. Dhandra, M. Hangarge, R. Hegadi, and V. S. Malemath, "Word level script identification in bilingual documents through discriminating features," in *Proc. ICSCN*, Chennai, India, Feb. 2007, pp. 630–635.

[32] B. V. Dhandra, P. Nagabhushan, M. Hangarge, R. Hegadi, and V. S. Malemath, "Script identification based on morphological reconstruction in document images," in *Proc. ICPR*, Hong Kong, Aug. 2006, pp. 950–953.

[33] D. Dhanya, A. G. Ramakrishnan, and P. B. Pati, "Script identification in printed bilingual documents," *Sadhana*, vol. 27, no. 1, pp. 73–82, Feb. 2002.

[34] J. Ding, L. Lam, and C. Y. Suen, "Classification of oriental and European scripts by using characteristic features," in *Proc. ICDAR*, Ulm, Germany, Aug. 1997, pp. 1023–1027.

[35] C. Djeddi, A. Gattal, L. Souici-Meslati, I. Siddiqi, Y. Chibani, and H. El Abed, "LAMIS-MSHD: A multi-script offline handwriting database," in *Proc. ICFHR*, Sep. 2014, pp. 93–97.

[36] M. A. Ferrer, A. Morales, and U. Pal, "LBP based line-wise script identification," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013, pp. 369–373.

[37] M. A. Ferrer, A. Morales, N. Rodríguez, and U. Pal, "Multiple training—One test methodology for handwritten word-script identification," in *Proc. 14th ICFHR*, Sep. 2014, pp. 754–759.

[38] B. P. Gaikwad, R. R. Manza, and G. R. Manza, "Video scene segmentation to separate script," in *Proc. IACC*, Ghaziabad, India, Feb. 2013, pp. 1269–1274.

[39] D. Ghosh and A. P. Shivaprasad, "Handwritten script identification using possibilistic approach for cluster analysis," *J. Indian Inst. Sci.*, vol. 80, no. 3, p. 215, 2000.

[40] D. Ghosh, T. Dube, and A. P. Shivaprasad, "Script recognition—A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2142–2161, Dec. 2010.

[41] S. Ghosh and B. B. Chaudhuri, "Composite script identification and orientation detection for Indian text images," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 294–298.

[42] J. Gllavata and B. Freisleben, "Script recognition in images with complex backgrounds," in *Proc. ISSPIT*, Athens, Greece, Dec. 2005, pp. 589–594.

[43] M. Hangarge, K. C. Santosh, and R. Pardeshi, "Directional discrete cosine transform for handwritten script identification," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013, pp. 344–348.
[44] P. S. Hiremath, S. Shivashankar, J. D. Pujari, and V. Mouneswara, "Script Identification in a handwritten document image using texture features," in *Proc. IACC*, Patiala, India, Feb. 2010, pp. 110–114.
[45] P. S. Hiremath and S. Shivashankar, "Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1182–1189, 2008.
[46] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 176–181, Feb. 1997.
[47] J. Hochberg, K. Bowers, M. Cannon, and P. Kelly, "Script and language identification for handwritten document images," *Int. J. Document Anal. Recognit.*, vol. 2, no. 2, pp. 45–52, Dec. 1999.
[48] S. Jaeger, H. Ma, and D. Doermann, "Identifying script on word-level with informational confidence," in *Proc. ICDAR*, Aug. 2005, pp. 416–420.
[49] F. K. Jaiem, S. Kanoun, M. Khemakhem, H. El Abed, and J. Kardoun, "Database for Arabic printed text recognition research," in *Image Analysis and Processing-ICIAP* (Lecture Notes in Computer Science), vol. 8156. A. Petrosino, Ed. Berlin, Germany: Springer, 2013, pp. 251–259.
[50] S. Kanoun, A. Ennaji, Y. Lecourtier, and A. M. Alimi, "Script and nature differentiation for Arabic and Latin text images," in *Proc. IWFHR*, Aug. 2002, pp. 309–313.
[51] M. Khoddami and A. Behrad, "Farsi and Latin script identification using curvature scale space features," in *Proc. NEUREL*, Belgrade, Serbia, Sep. 2010, pp. 213–217.
[52] L. Lam, J. Ding, and C. Y. Suen, "Differentiating between oriental and european scripts by statistical features," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 12, no. 1, pp. 63–79, 1998.
[53] D. S. Lee, C. R. Nohl, and H. S. Baird, "Language identification in complex, unoriented, and degraded document images," in *Proc. DAS*, 1996, pp. 76–98.
[54] L. Li and C. L. Tan, "Script identification of camera-based images," in *Proc. ICPR*, Tampa, FL, USA, Dec. 2008, pp. 1–4.
[55] X.-R. Lin, C.-Y. Guo, and F. Chang, "Classifying textual components of bilingual documents with decision-tree support vector machines," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 498–502.
[56] S. Lu and C.-L. Tan, "Automatic detection of document script and orientation," in *Proc. ICDAR*, Parana, Sep. 2007, pp. 237–241.
[57] S. Lu and C. L. Tan, "Script and language identification in noisy and degraded document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 14–24, Jun. 2008.
[58] H. Luqman, S. A. Mahmoud, and S. Awaida, "KAFD Arabic font database," *Pattern Recognit.*, vol. 47, no. 6, pp. 2231–2240, Jun. 2014.
[59] H. Ma and D. Doermann, "Gabor filter based multi-class classifier for scanned document images," in *Proc. ICDAR*, 2003, pp. 968–972.
[60] S. A. Mahmoud *et al.*, "KHATT: Arabic offline handwritten text database," in *Proc. 13th ICFHR*, Sep. 2012, pp. 447–452.
[61] U. Marti and H. Bunke, "The IAM-database: An English sentence database for offline handwriting recognition," *Int. J. Document Anal. Recognit.*, vol. 5, no. 1, pp. 39–46, 2002.
[62] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," *Proc. IEEE*, vol. 80, no. 7, pp. 1029–1058, Jul. 1992.
[63] P. Nagabhushan, S. A. Angadi, and B. S. Anami, "An intelligent pin code script identification methodology based on texture analysis using modified invariant moments," in *Proc. ICCR*, 2005, pp. 615–623.
[64] A. M. Namboodiri and A. K. Jain, "On-line Script Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 124–130, Jan. 2004.
[65] S. M. Obaidullah, C. Halder, N. Das, and K. Roy, "Numeral script identification from handwritten document images," in *Proc. 11th IMCIP*, 2015, pp. 585–594.
[66] S. M. Obaidullah, R. Karim, S. Shaikh, C. Halder, N. Das, and K. Roy, "Transform based approach for indic script identification from handwritten document images," in *Proc. 3rd ICSCN*, Mar. 2015, pp. 1–7.
[67] S. M. Obaidullah, K. Roy, and N. Das, "Comparison of different classifiers for script identification from handwritten document," in *Proc. ISPCC*, Sep. 2013, pp. 1–6.
[68] M. C. Padma and P. A. Vijaya, "Monothetic separation of Telugu, Hindi and English text lines from a multi script document," in *Proc. SMC*, Oct. 2009, pp. 4870–4875.
[69] M. C. Padma and P. A. Vijaya, "Entropy based texture features useful for automatic script identification," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 115–120, 2010.
[70] U. Pal and B. B. Chaudhuri, "Automatic separation of words in multilingual multi-script Indian documents," in *Proc. ICDAR*, Ulm, Germany, Aug. 1997, pp. 576–579.
[71] U. Pal and B. B. Chaudhuri, "Script line separation from Indian multi-Script documents," in *Proc. ICDAR*, Bangalore, India, 1999, pp. 406–409.
[72] U. Pal and B. B. Chaudhuri, "Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line," in *Proc. ICDAR*, Seattle, WA, USA, Sep. 2001, pp. 790–794.
[73] U. Pal, S. Sinha, and B. B. Chaudhuri, "Multi-script line identification from Indian documents," in *Proc. ICDAR*, Bangalore, India, 2003, pp. 880–884.
[74] U. Pal and A. Sarkar, "Recognition of printed urdu script," in *Proc. ICDAR*, Bangalore, India, 2003, pp. 1183–1187.
[75] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition of six popular Indian scripts," in *Proc. ICDAR*, Parana, Sep. 2007, pp. 749–753.
[76] U. Pal and B. B. Chaudhuri, "Identification of different script lines from multi-script documents," *Image Vis. Comput.*, vol. 20, nos. 13–14, pp. 945–954, Dec. 2002.
[77] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," *Pattern Recognit.*, vol. 37, no. 9, pp. 1887–1899, Sep. 2004.
[78] U. Pal, "Automatic script identification: A survay," *J. VIVEK*, Bombay, vol. 16, no. 3, pp. 26–35, 2006.
[79] J. Pan and Y. Tang, "A rotation-robust script identification based on BEMD and LBP," in *Proc. ICWAPR*, Guilin, China, Jul. 2011, pp. 165–170.
[80] W. M. Pan, C. Y. Suen, and T. D. Bui, "Script identification using steerable Gabor filters," in *Proc. ICDAR*, Aug. 2005, pp. 883–887.
[81] P. B. Pati and A. G. Ramakrishnan, "HVS inspired system for script identification in Indian multi-script documents," in *Document Analysis Systems VII. DAS*, (Lecture Notes in Computer Science), vol. 3872. H. Bunke and A. L. Spitz, Eds. Berlin, Germany: Springer, 2006, pp. 380–389.
[82] P. B. Pati, S. SabariRaju, N. Pati, and A. G. Ramakrishnan, "Gabor filters for document analysis in Indian bilingual documents," in *Proc. ICISIP*, 2004, pp. 123–126.
[83] P. B. Pati and A. G. Ramakrishnan, "Word level multi-script identification," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1218–1229, 2008.
[84] S. B. Patil and N. V. Subbareddy, "Neural network based system for script identification in Indian documents," *Sadhana*, vol. 27, no. 1, pp. 83–97, 2002.
[85] G. S. Peake and T. N. Tan, "Script and language identification fromdocument images," in *Computer Vision-ACCV* (Lecture Notes in Computer Science), vol. 1352. R. Chin and T. C. Pong, Eds. Berlin, Germany: Springer, 1998, pp. 97–104.
[86] L. Peng, C. Liu, X. Ding, and H. Wang, "Multilingual document recognition research and its application in China," in *Proc. DIAL*, 2006, pp. 126–132.
[87] T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu, and C. L. Tan, "Video script identification based on text lines," in *Proc. ICDAR*, 2011, pp. 1240–1244.
[88] B. Philip and R. D. S. Samuel, "A novel bilingual OCR for printed Malayalam-English text based on Gabor features and dominant singular values," in *Proc. DIP*, Bangkok, Thailand, Mar. 2009, pp. 361–365.
[89] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification—The state of the Art," *Pattern Recognit.*, vol. 22, no. 2, pp. 107–131, 1989.
[90] K. A. Prakash, G. Rajesh, U. A. Dinesh, M. Krisnamoorthi, and N. V. Subbareddy, "Text line script identification for a trilingual document," in *Proc. ICCCNT*, 2010, pp. 1–3.
[91] R. Rani, R. Dhir, and G. S. Lehal, "Script identification of pre-segmented multi-font characters and digits," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013, pp. 1150–1154.
[92] R. Rani, R. Dhir, and G. S. Lehal, "Performance analysis of feature extractors and classifiers for script recognition of English and Gurmukhi words," in *Proc. DAR*, 2012, pp. 30–36.
[93] M. I. Razzak, S. A. Hussain, and M. Sher, "Numeral recognition for Urdu script in unconstrained environment," in *Proc. ICET*, Oct. 2009, pp. 44–47.
[94] H. Rezaee, M. Geravanchizadeh, and F. Razzazi, "Automatic language identification of bilingual english and farsi scripts," in *Proc. AICT*, Baku, Azerbaijan, Oct. 2009, pp. 1–4.

[95] K. Roy, A. Banerjee, and U. Pal, "A system for word-wise handwritten script identification for Indian postal automation," in *Proc. INDICON*, Dec. 2004, pp. 266–271.
[96] K. Roy, U. Pal, and B. B. Chaudhuri, "Neural network based word-wise handwritten script identification system for Indian postal automation," in *Proc. ICISIP*, Jan. 2005, pp. 240–245.
[97] K. Roy and K. Majumder, "Trilingual script separation of handwritten postal document," in *Proc. ICVGIP*, Dec. 2008, pp. 693–700.
[98] K. Roy, S. K. Das, and S. M. Obaidullah, "Script identification from handwritten document," in *Proc. NCVPRIPG*, Dec. 2011, pp. 66–69.
[99] K. Roy, A. Alaei, and U. Pal, "Word-wise handwritten persian and roman script identification," in *Proc. ICFHR*, Kolkata, India, Nov. 2010, pp. 628–633.
[100] A. Saïdani, A. Kacem, and A. Belaïd, "Co-occurrence matrix of oriented gradients for word script and nature identification," in *Proc. 13th ICDAR*, Aug. 2015, pp. 16–20.
[101] A. Saïdani, A. K. Echi, and A. Belaïd, "Identification of machine-printed and handwritten words in Arabic and Latin Scripts," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013, pp. 798–802.
[102] R. S. Kunte and R. D. S. Samuel, "On separation of kannada and english words from a bilingual document employing gabor features and radial basis function neural network," in *Proc. ICCR*, 2005, pp. 640–644.
[103] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "Word level script identification from bangla and devanagri handwritten texts mixed with roman script," *J. Comput.*, vol. 2, no. 2, pp. 103–108, 2010.
[104] N. Sharma, S. Chanda, U. Pal, and M. Blumenstein, "Word-wise script identification from video frames," in *Proc. ICDAR*, Washington, DC, USA, Aug. 2013, pp. 867–871.
[105] N. Sharma, R. Mandal, R. Sharma, P. P. Roy, U. Pal, and M. Blumenstein, "Multi-lingual text recognition from video frames," in *Proc. 13th ICDAR*, Aug. 2015, pp. 951–955.
[106] N. Sharma, R. Mandal, R. Sharma, U. Pal, and M. Blumenstein, "ICDAR2015 competition on video script identification (CVSI 2015)," in *Proc. 13th ICDAR*, Aug. 2015, pp. 1196–1200.
[107] N. Sharma, U. Pal, and M. Blumenstein, "A study on word-level multi-script identification from video frames," in *Proc. IJCNN*, Beijing, China, Jul. 2014, pp. 1827–1833.
[108] N. Sharma, R. Mandal, R. Sharma, U. Pal, and M. Blumenstein, "Bag-of-visual words for word-wise video script identification: A study," in *Proc. IJCNN*, Killarney, Ireland, Jul. 2015, pp. 1–7.
[109] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. ICDAR*, Sep. 2011, pp. 1485–1489.
[110] B. Shi, C. Yao, C. Zhang, X. Guo, F. Huang, and X. Bai, "Automatic script identification in the wild," in *Proc.13th ICDAR*, Aug. 2015, pp. 531–535.
[111] B. Shi, X. Bai, and C. Yao, "Script identification in the wild via discriminative convolutional neural network," *Pattern Recognit.*, vol. 52, pp. 448–458, Apr. 2016.
[112] P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein, and C. L. Tan, "Gradient-angular-features for word-wise video script identification," in *Proc. ICPR*, Stockholm, Sweden, Aug. 2014, pp. 3098–3103.
[113] P. Shivakumara, Z. Yuan , D. Zhao , T. Lu , and C. L. Tan, "New gradient-spatial-structural features for video script identification," *Comput. Vis. Image Understand.*, vol. 130, pp. 35–53, Dec. 2015.
[114] P. K. Singh, R. Sarkar, and M. Nasipuri, "Offline script identification from multilingual indic-script documents: A state-of-the-art," *Comput. Sci. Rev.*, vols. 15–16, pp. 1–28, Aug. 2015.
[115] P. K. Singh, R. Sarkar, M. Nasipuri, and D. Doermann, "Word-level script identification for handwritten indic scripts," in *Proc. 13th ICDAR*, Aug. 2015, pp. 1106–1110.
[116] V. Singhal, N. Navin, and D. Ghosh, "Script-based classification of hand-written text documents in a multilingual environment," in *Proc. IWRIDEMIM*, Mar. 2003, pp. 47–54.
[117] F. Slimane, R. Ingold, S. Kanoun, M. A. Alimi, and J. Hennebert, "A new arabic printed text image database and evaluation protocols," in *Proc. ICDAR*, Jul. 2009, pp. 946–950.
[118] A. Spitz, "Script and language determination from document images," in *Proc. Symp. Document Anal. Inf. Retr.*, Las Vegas, NV, USA, Apr. 1994, pp. 229–235.
[119] A. L. Spitz, "Determination of the script and language content of document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 3, pp. 235–245, Mar. 1997.
[120] G. X. Tan, C. Viard-Gaudin, and A. C. Kot, "Information retrieval model for online handwritten script identification," in *Proc. ICDAR*, Jul. 2009, pp. 336–340.
[121] T. N. Tan, "Rotation invariant texture features and their use in automatic script identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 751–756, Jul. 1998.
[122] Y. Tho and Y. Y. Tang, "Discrimination of oriental and euramerican scripts using fractal feature," in *Proc. ICDAR*, Seattle, WA, USA, Sep. 2001, pp. 1115–1119.
[123] T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu, and C. L. Tan, "Video Script Identification based on Text Lines," in *Proc. ICDAR*, Beijing, China, Sep. 2011, pp. 1240–1244.
[124] A. Ul-Hasan, M. Z. Afzal, F. Shafait, M. Liwicki, and T. M. Breuel, "A sequence learning approach for multiple script identification," in *Proc. 13th ICDAR*, Aug. 2015, pp. 1046–1050.
[125] T. N. Vikram and D. S. Guru, "Appearance based models in document script identification," in *Proc. ICDAR*, Parana, Sep. 2007, pp. 709–713.
[126] B. Waked, S. Bergler, C. Y. Suen, and S. Khoury, "Skew detection, page segmentation, and script classification of printed document images," in *Proc. ICSMC*, San Diego, CA, USA, Oct. 1998, pp. 4470–4475.
[127] N. Wang, L. Lam, and C. Y. Suen, "Noise tolerant script identification of printed oriental and English documents using a downgraded pixel density feature," in *Proc. ICPR*, Istanbul, Turkey, Aug. 2010, pp. 2037–2040.
[128] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. ECCV*, 2010, pp. 591–604.
[129] S. L. Wood, X. Yao, K. Krishnamurthi, and L. Dang, "Language identification for printed text independent of segmentation," in *Proc. ICIP*, vol. 3. Washington, DC, USA, Oct. 1995, pp. 428–431.
[130] S. Wshah, G. Kumar, and V. Govindaraju, "Statistical script independent word spotting in offline handwritten documents," *Pattern Recognit.*, vol. 47, no. 3, pp. 1039–1050, Mar. 2014.
[131] L. Zeng, Y.-Y. Tang, and T.-H. Cheni, "Multi-scale wavelet texture-based scrip identification method," *Chin. J. Comput.*, vol. 23, no. 7, pp. 699–704, 2000.
[132] D. Zhao, P. Shivakumara, S. Lu, and C. L. Tan, "New spatial-gradient-features for video script identification," in *Proc. DAS*, Mar. 2012, pp. 38–42.
[133] L. Zhou, Y. Lu, and C. L. Tan, "Bangla/English script identification based on analysis of connected component profiles," in *Document Analysis Systems VII. DAS* (Lecture Notes in Computer Science), vol. 3872. H. Bunke and A. L. Spitz, Eds. Berlin, Germany: Springer, 2006, pp. 243–254.
[134] L. Zhou, X. J. Ping, E. G. Zheng, and L. Guo, "Script identification based on wavelet energy histogram moment features," in *Proc. ICSP*, Beijing, China, Oct. 2010, pp. 980–983.

KURBAN UBUL (M'15) has been a Researcher with the Xinjiang Laboratory of Multi-language Information Technology since 2000. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University, China. He has authored three books, one book chapter, and over 60 papers, presented at numerous international conferences. His research interests include image processing, pattern recognition, speech signal processing, digital signal processing, and education technology. He is a member of the International Association of Pattern Recognition, the IEEE Computer Society, the Association for Computing Machinery, and the China Computer Federation. He served as the technical committee member/reviewer of many international conferences such as the CCBR2017, the IGS2017, the CCBR2016, the CSA2016, the Isca2015, the ICDAR2015, the ICMIAC2014, the WOSSPA2013, the ISIEA2012, and the ICSP2012. He is a Committee Member of the Pattern Recognition Professional Committee for China Artificial Intelligence Association and the Pattern Recognition and Machine Intelligence Professional Committee for China Automation Association. He currently serves as an Editor of the *Singapore Journal of Scientific Research*, the *Cyber Journals*, the *Image and Signal Processing*, and the *Journal of Security and Safety Technology*. He was a Reviewer of the IEEE IHMS, the IET Biometrics, and the IJITM.

GULZIRA TURSUN received the B.E. degree in communication engineering and the M.S. degree in signal and information processing from the Xinjiang University, China, in 2013 and 2016, respectively. Her research is focused on image processing, pattern recognition, and document identification.

ALIMJAN AYSA is currently an Associate Professor with the Network and Information Technology Center, Xinjiang University. He has also been as a Researcher with the Xinjiang Laboratory of Multi-language Information Technology since 2002. His research interests include natural language processing, pattern recognition, and digital signal processing. He has authored one book and over 30 papers. His work has appeared in many Chinese core journals such as the *Computer Engineering and Application*, the *Chinese Information Processing*, and *Computer Application*. He is a member of the IEEE Computer Society, and the China Computer Federation.

DONATO IMPEDOVO (M'08) received the M.Eng. degree (*cum laude*) in Computer engineering and the Ph.D. degree in computer engineering. He is currently an Associate Professor with the Department of Computer Science, University of Bari. He has co-authored over 70 articles on these fields in both international journals and conference proceedings. He is also involved in research transfer activities and industrial research. He has managed over 20 projects funded by public institutions and by private SMEs. His research interests are in the field of pattern recognition, signal processing, and biometrics. He received the Distinction Award in 2009 from the International Conference on Computer Recognition Systems (endorsed by IAPR), and the First Prize of the first Nereus-Euroavia Academic Competition on GMES in 2012. He is a Reviewer of many international journals including the IEEE Access, the IEEE T-SMC, the *Pattern Recognition*, and many others. He was the General Co-Chair of the International Workshop on Emergent Aspects in Handwritten Signature Processing and the International Workshop on Image-Based Smart City Application. He is an IAPR member.

GIUSEPPE PIRLO (M'92–SM'13) received the degree in computer science (*cum laude*) from the Department of Computer Science, University of Bari, Italy, in 1986. Since 1986, he has been carrying out research in the field of computer science and neuroscience, signal processing, handwriting processing, automatic signature verification, biometrics, pattern recognition and statistical data processing. Since 1991, he has been an Assistant Professor with the Department of Computer Science, University of Bari, where he is currently an Associate Professor.

He developed several scientific projects and authored over 250 papers on international journals, scientific books and proceedings.

Prof. Pirlo is currently an Associate Editor of the IEEE Transactions on Human–Machine Systems. He also serves as a Reviewer for many international journals including the IEEE T-PAMI, the IEEE T-FS, the IEEE T-SMC, the IEEE T-EC, the IEEE T-IP, the IEEE T-IFS, the *Pattern Recognition*, the IJDAR, and the IPL. He was the General Chair of the International Workshop on Emerging Aspects in Handwriting Signature Processing, Naples, in 2013, the International Workshop on Image-based Smart City Applications, Genoa, in 2015, and the General Co-Chair of the International Conference on Frontiers in Handwriting Recognition, Bari, in 2012. He was a reviewer in the scientific committee and program committee of many international conferences in the field of computer science, pattern recognition and signal processing, such as the ICPR, the ICDAR, the ICFHR, the IWFHR, the ICIAP, the VECIMS, and the CISMA.

He is also the editor of several books. He was an Editor of the Special Issue Handwriting Recognition and Other PR Applications of the *Pattern Recognition Journal* in 2014 and the Special Issue Handwriting Biometrics of the *IET Biometrics Journal* in 2014. He was the Guest Editor of the Special Issue of the *Je-LKS Journal* of the e-Learning and Knowledge Society Steps toward the Digital Agenda: Open Data to Open Knowledge in 2014. He is currently the Guest Co-Editor of the Special Issue of the IEEE Transactions on Human–Machine Systems on Drawing and Handwriting Processing for User-Centered Systems.

Prof. Pirlo is a member of the Governing Board of Consorzio Interuniversitario Nazionale per l'Informatica (CINI), a member of the Governing Board of the Società Italiana di e-Learning and the e-learning Committee of the University of Bari. He is currently the Deputy Representative of the University of Bari in the Governing Board of CINI. He is also the Managing Advisor of the University of Bari for the Digital Agenda and Smart Cities. He is the Chair of the Associazione Italiana Calcolo Automatico-Puglia. He is also a member of the Gruppo Italiano Ricercatori Pattern Recognition, the International Association Pattern Recognition, the Stati Generali dell'Innovazione, and the Gruppo Ingegneria Informatica.

TUERGEN YIBULAYIN is currently a Professor and a Ph.D. supervisor with the School of Information Science and Engineering, Xinjiang University, China, and the Vice-Director of the Xinjiang Laboratory of Multi-language Information Technology, China. His research is focused on natural language processing, machine translation, information retrieval, and pattern recognition. He has authored over 100 papers. He has won many science and research awards including the Second National Prize for Progress in Science and Technology of China. He currently serves as the Standing Director of the Chinese Information Society of China, the Deputy Director of the Professional Committee of Ethnic Minorities in Chinese Information Society, and an Editor of the *Journal of Chinese Information Processing*.