Language, Script, and Font Recognition

9

Umapada Pal and Niladri Sekhar Dash

Contents

Introduction 292
Language Identification 294
Language Overview 294
Origin of Language 296
Difficulties in Language Identification 300
Existing Approaches for Language Identification 301
Script Identification 302
Script Overview 302
Single- and Multiscript Documents 305
Script Identification Technology and Challenges 307
Machine-Printed Script Identification 313
Handwritten Script Identification 316
Font and Style Recognition 317
Font Terminology 322
Font Generation 322
Font Variation 322
Recognition Strategies for Font and Style 323
Conclusions 325
Cross-References 326
Notes and Comments 327
References 327
Further Reading 330

U. Pal (✉)
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
e-mail: umapada@isical.ac.in

N.S. Dash
Linguistic Research Unit, Indian Statistical Institute, Kolkata, India
e-mail: niladri@isical.ac.in

D. Doermann, K. Tombre (eds.), *Handbook of Document Image Processing and Recognition*, DOI 10.1007/978-0-85729-859-1_9,

Abstract

Automatic identification of a language within a text document containing multiple scripts and fonts is a challenging task, as it is not only linked with the shape, size, and style of the characters and symbols used in the formation of the text but also admixed with more crucial factors such as the forms and size of pages, layout of written text, spacing between text lines, design of characters, density of information, directionality of text composition, etc. Therefore, successful management of the various types of information in the act of character, script, and language recognition requires an intelligent system that can elegantly deal with all these factors and issues along with other secondary factors such as language identity, writing system, ethnicity, anthropology, etc. Due to such complexities, identification of script vis-à-vis language has been a real challenge in optical character recognition (OCR) and information retrieval technology. Considering the global upsurge of so-called minor and/or unknown languages, it has become a technological challenge to develop automatic or semiautomatic systems that can identify a language vis-à-vis a script in which a particular piece of text document is composed. Bearing these issues in mind, an attempt is initiated in this chapter to address some of the methods and approaches developed so far for language, script, and font recognition for written text documents. The first section, after presenting a general overview of language, deals with the information about the origin of language, the difficulties faced in language identification, and the existing approaches to language identification. The second section presents an overview of script, differentiates between single- and multiscript documents, describes script identification technologies and the challenges involved therein, focuses on the process of machine-printed script identification, and then addresses the issues involved in handwritten script identification. The third section tries to define font terminologies, addresses the problems involved in font generation, refers to the phenomenon of font variation in a language, and discusses strategies for font and style recognition. Thus, the chapter depicts a panoramic portrait of the three basic components involved in OCR technology: the problems and issues involved, the milestones achieved so far, and the challenges that still lie ahead.

Keywords

Document image analysis • Natural language processing • OCR technology • Language and script identification • Font and style recognition • Multiscript OCR • Handwriting recognition • Orthography • Grapheme • Diacritic • Allograph

Introduction

The science that tries to understand language and its properties from the perspective of human cognition and communication considers script and font as two valuable components. Interpretation and analysis of script and font provide necessary insight

to understand how people take advantage of these elements to fabricate a network of information interchange through encoding knowledge that is sharable across the members of communities in diachronic and synchronic dimensions. In this frame, both script and font are important avenues via which language achieves continuation of its existence and growth through all spatiotemporal barriers across generations. On the axis of empirical science, understanding script and font is a tiny step towards understanding the more complex process of encoding knowledge and information within a collectively approved set of symbols that stand as an approximate visual representative of the sounds used in the spoken form of a language. Also, understanding the form and function of script and font becomes relevant in the wider frame of language policy and language planning, where these two elements work as crucial factors in language survival, language growth, mass literacy, and grass-roots education.

On the other hand, the technology that tries to understand language from the perspective of machine learning and knowledge representation considers script and font as two valuable pools of information whose systematic interpretation can make a computer system more robust and penetrative in extraction of information embedded within written texts formed from scripts with different fonts, styles, and designs. Therefore, it becomes necessary for technologies such as optical character recognition (OCR) to understand in minute detail the fine-grained aspects linked with scripts and fonts of languages, since better understanding of the features of scripts and fonts increases the ability of a system to interpret and extract information required for developing the technology.

Keeping both of these missions in view, in this chapter an attempt is made to address language identification, script identification, and font-cum-style recognition, which are three important areas of OCR technology. These are also important problems in the human cognition processes as people fail to identify a language, script or font if they are not previously trained with necessary information regarding these elements. This chapter, however, does not address these issues, since human recognition of language, script, and font is more directly linked with human psychology and the cognition process rather than with machine learning and technology development. Therefore, the primary goal of this chapter is to present a short yet insightful survey on the problems and issues relating to the development of tools and technology for machine recognition of language, script, and font of a document as well as to refer to the achievements made in these domains. This chapter also aims to provide some glimpses on the present state of the art of this technology and to draw the attention of readers towards new directions in this scientific endeavor.

The chapter is broadly divided into three main sections, covering language identification (section "Language Identification"), script identification (section "Script Identification"), and font and style recognition (section "Font and Style Recognition"). Each section has several subsections where specific problems in each broad area are adequately addressed. An overview of language is provided in section "Language Overview," the origin of language in section "Origin of Language," difficulties faced in language identification in section "Difficulties

in Language Identification," and existing approaches to language identification in section "Existing Approaches for Language Identification." The third section further subdivided into "Script Overview", "Single- and Multiscript Documents", "Script Identification Technology and Challenges", "Machine-Printed Script Identification", and "Handwritten Script Identification". The fourth section defines font terminology (section "Font Terminology"), problems of font generation (section "Font Generation"), font variation (section "Font Variation"), and font and style recognition strategies (section "Recognition Strategies for Font and Style"). Thus, this chapter presents a general overview on the three basic components involved in document analysis technology. From this chapter, the reader will primarily learn about three important aspects of document analysis technology: the problems and issues involved, the milestones achieved so far, and the challenges that still remain ahead.

Language Identification

Language Overview

The general use of the term *language* is embedded with different shades of meaning, which a trained person can decipher if (s)he seriously takes into account the finer sense variations invoked by the term. In general, the term refers to the act of exchange of verbal linguistic signals (known as speech) within a given situation with fellow interlocutors on some topics of social relevance in a particular social setting. Thus, the act of speaking becomes a social-cognitive event that activates the *linguistic* and *communicative competence* as well as *speech repertoire* of the speakers involved in the spoken interaction. On the individual scale, this particular linguistic event of communication and information interchange becomes an *idiolect* when the inherent implication of the general term is minimized to refer to the use of the system by an individual at a particular point in time and place in different sociocultural settings.

The term *language* can also be used to refer to a particular variety of speech or writing produced by the members of a particular social group; For instance, set expressions and phrases such as *scientific language*, *technical language*, *media language*, *corporate language*, *business language*, *adult language*, *child language*, *woman language*, *secret language*, *underworld language*, etc. actually refer to specific varieties of language formed and used by different social groups in specific sociocultural contexts to address specific functional needs.

In the area of language acquisition and language teaching, on the other hand, phrases such as *first language*, *second language*, *third language*, *mother language*, *foreign language*, etc. refer to an abstract system of linguistic properties and communication that combines the collective totality of both speech and writing behaviors of the members of the speech community. It also refers to the innate knowledge of the system by the members of the community.

On the scale of time, the term *language* is used in both synchronic and diachronic senses. In the synchronic sense, it refers to the language used at a particular point in time at one or many places, such as *Modern English*, *Modern French* or *Modern Spanish*, whereas in the diachronic sense, it refers to the language as found to be used at various points of time in the history of its birth and growth, such as *Old English*, *Medieval English*, *Victorian English*, *Modern English*, etc.

In the realm of society and culture, the term *language* is used to classify and cluster groups of different speech varieties, such as *pidgins* and *creoles*, that have unique linguistic identities in relation to the standard varieties used by the members of a society.

All the examples and varieties mentioned above invariably fall under the generic term *natural language*, which is often contrasted with those artificially constructed systems that are used to expound one or more conceptual areas, such as *mathematical language*, *computer language*, *formal language*, *logical language*, etc. This contrast between natural and artificial may be further expanded to include the uniqueness of languages such as *Volapuk* and *Esperanto*, which are artificially devised for providing a universally approved platform for the purpose of communication among people across the world.

At an abstract level, the term *language* may be postulated as a bundle of features of human behavior – the universal properties that are present in all human speeches and writing systems but strikingly missing from animal communication systems. In this sense, the term may be characterized by some *design features*, such as *productivity*, *prevarication*, *duality of patterning*, *learnability*, etc., as proposed by linguists [1, 2]. In the purest abstract sense, however, the term *language* refers to the *innate biological faculty* of an individual through which a human being is empowered to learn and use natural language(s) in all possible ways. This sense is implicit within the concept known as language acquisition device (LAD) – an area of intensive research in *biolinguistics*, *psycholinguistics*, *language acquisition*, and *cognitive linguistics* [4].

We can learn about the language used by a speech community by studying the varieties or dialects used by the members of that community as well as by developing a realistic policy concerning the selection and use of different varieties at different geocultural locations. This is the area of geolinguistics and ecolinguistics, where the term *language* acquires a unique identity to refer to the dialects and other regional varieties endowed with unique geocultural environments and natural elements.

In a similar fashion, the term *language* may enter into the sphere of technical intricacies when we talk about issues such as *language teaching*, *language learning*, *language planning*, and *language policy*. In these areas, we require adequate knowledge to understand the *first language* (i.e., mother tongue), which is distinguishable from the *second language* (i.e., a language other than one's mother tongue) used for several practical purposes, such as *government work*, *education*, *administration*, *migration*, *rehabilitation*, *political campaigns*, *tours and travel*, *commercial activities*, etc.

Since the term *language* is polysemous, it is used to refer to not only human languages but also a variety of other systems of communication – both human and nonhuman – which are natural but not *language* in the true sense of the term; For instance, let us consider phrases such as *sign language*, *body language* or *animal language*. One must agree that the term *language* is used in these cases in a highly metaphorical or figurative sense. These are termed so because these communication systems, in spite of their unique functional identities, share some attributes which are common to the features of natural languages.

Several languages use two different words to translate the English word *language*; For instance, French uses *langage* and *langue*, Italian uses *linguaggio* and *lingua*, Spanish uses *lenguaje* and *lengua*, etc. In each of these languages, the difference between the two words correlates with the difference between the two senses evoked by the English word; For example, in French, while *langage* refers to the language in general, *langue* refers to particular languages. This happens because, in English, it is assumed that a human being not only possesses a *language* (e.g., English, Chinese, German, Bengali, Hindi, etc.) but also has *the language* (i.e., the language faculty) due to which a human being is able to communicate with fellow members. Although it is known that possession of the *language faculty* clearly distinguishes human beings from animals [2], the important fact to note is that one cannot possess (or use) a natural language unless one possesses (or uses) some particular language.

The discussion above shows that the question "What is language?" actually carries with it a presupposition that each natural language spoken in the world is a specific instance of something more general – a unique communication system gifted to the human race. Since linguists are concerned primarily with natural languages, they want to know whether all natural languages have something in common, not being shared by other, human or nonhuman communication systems, so that it is possible to apply to each of them the term *language* and deny the use of that term to other systems of communication. This implies that we need to verify whether it is possible to assign the characteristic features of human languages to other systems of communication before we can call them *language*.

Origin of Language

The debate about the origin of language has continued for many years, with scholars debating whether it is possible to account for natural language by using only the basic mechanisms of learning or whether one needs to postulate some special built-in language devices for this purpose. Scholars such as Skinner [3], who believed that human language was nothing but a *learning-only mechanism*, argued that childhood conditioning or modeling can account for the complexities involved in a natural language. On the other hand, Chomsky [4], Pinker and Bloom [5], and Pinker [6], who support the Cartesian and Darwinian models of language acquisition and learning, believe that the phenomenon of the ease and speed with which a human

child learns a natural language requires something more than a mere behavioristic model to understand how human language originated or evolved.

From analysis of the brain, it is observed that, in the case of mammals except for human beings, both hemispheres look very much alike. In the case of human brains, the hemispheres are different in shape and function. It is assumed that, in the prehistoric period, due to continuous struggles (both physical and cerebral) of human beings with hostile Nature, one of the hemispheres of the human brain developed with reduced capacity. As a result, the neural network in the human brain, instead of spreading in all directions, mostly expanded in linear order within the brain. Due to this development, the left hemisphere is hardly successful to relate things in a normal full-blown multidimensional scheme. However, this feature becomes an advantage as the reduced capability of the left hemisphere proves to be highly useful for ordering things at a linear level. And that is exactly what a human language needs – the capacity of the brain for linear arrangement of linguistic elements. Human language needs the ability of the brain to convert full-blown multidimensional natural events to be reproduced in linear sequences of sounds, and vice versa.

While it is certain that human speech developed long before the advent of writing systems (as it is observed that many human societies have speech but no writing system), no one knows for certain how human languages originated or evolved.

The question of all ages and civilizations is: When and how did language emerge? Did it begin at the time when *Homo sapiens* came into the world nearly 4–5 million years ago? Or did it start with the evolution of *Modern Man* (i.e., Cro-Magnon) nearly 125,000 years ago? We are not sure about the actual antiquity of the origin of language, as there is no evidence to justify our assumptions. Similarly, we are not sure if Neanderthal Man could speak. Although he had a brain that was larger than normal human beings of the Modern Age, his voice box was positioned much higher in his throat (like that of apes), and this position is not very convenient for producing fluent speech. The assumption, therefore, is that it is highly unlikely that Neanderthal Man could speak, although this has been an intriguing question for centuries. Even then, the following question still lingers: How did humans obtain *language* which is so different from the nonverbal signals produced by humans (e.g., *gesture*, *facial expression*, *kinesis*, *proxemics*, *body movement*, *smiling*, *posture*, etc.) as well as from the verbal signals (e.g., *barking*, *roaring*, *howling*, *growling*, *calling*, *chirping*, *twittering*, etc.) produced by animals?

Animals often make use of various physical signals and gestures, which represent their specific biological needs and responses. They, however, cannot produce and use symbols as a human being does in an arbitrary fashion following the conventions of the speech community to which the individual belongs. In the animal world, a particular physical signal or posture usually refers to a particular piece of information, and the interface between the signal and the information is always iconic across time and space. The feature of multisemanticity, which is one of the most important attributes of human language, is lacking from "animal language," and therefore, animal language fails to provide necessary clues and insights to trace the origin of human language. Human language, as a codified system of

symbols, is framed with several layers through taxonomic organization of linguistic properties such as *phoneme*, *morpheme*, *word*, *sentence*, *meaning*, *text*, and *discourse*.

The oversimplified history of the evolution of human civilization postulates that *Homo sapiens* evolved, as a subdivision, from the hominoid family through the following stages:

(a) Humans split from the apes nearly 3 million years ago.
(b) The tool-using *Homo habilis* (i.e., handy man) emerged nearly 2 million years ago.
(c) *Homo erectus* (i.e., upright man) came into existence nearly 1.5 million years ago. He was able to use fire.
(d) Archaic *Homo sapiens* (i.e., archaic wise man) arrived nearly 300,000 years ago.
(e) *Homo sapiens* (i.e., modern human) came into being nearly 200,000 years ago.
(f) *Homo sapiens* started using stone and other raw materials such as bone and clay nearly 50,000 years ago.
(g) Around this time, *Homo sapiens* started carving and graving on cave walls. This may be assumed as a carnal stage of communication that later evolved into language.

It is assumed that the characteristic features of grammar of a natural language (such as the distinction in formation of a sentence and a phrase, the organization of inflection classes in paradigm, etc.) may provide necessary clues about the prehistory of a language. When the vocal tract of *Homo sapiens* was reshaped in the course of evolution, it provided a better platform for syllabically organized speech output. This perhaps made it possible to increase the vocabulary of humans, which, in return, helped to develop linear sequences of syllables to frame grammar vis-à-vis syntax with active participation of the neural mechanism that controlled syllable structure. Analysis of sentences of natural languages reveals several features based on which it makes sense to assume syntax as a byproduct of characteristics of syllables (e.g., grammatical *subjects* may be the byproducts of onset margins of speech). This hypothesis comes closer to evidence acquired from biological anthropology, ape language studies, and brain neurophysiology [6].

For generations, one of the primary questions of mainstream linguistics has been how language came into being. Speculations to reply to this question are myriad. Some people considered that human language was invented by our earliest ancestors, who had genetic and physiological properties to develop complex sounds and organize these into meaningful strings to from larger constructions such as words and sentences. This theory is known as the monogenetic theory or monogenesis [6]. On the other hand, the polygenetic theory or polygenesis assumes that human language was invented many times by many peoples at different points in time. For this reason, it is possible to reconstruct the earlier forms of a language, but one cannot go far through the cycles of change before the meanders truly obliterate any possibility of reconstructing protoforms. Scholars suggest that one can, at best,

go back only 10,000 years or so. After that, the trail is lost in the river of oblivion. So, there is no chance for us to know what lay before this.

In the last century, linguists developed a method known as historical reconstruction, in which, based on linguistic evidence from modern languages, one can sufficiently reliably reconstruct the protoforms of a language not available today. Some scholars believed that, by applying this process, they would be able to trace the linguistic evidence to prehistoric periods. However, the primary limitation of this approach is that the method of hypothetical reconstruction can, at best, supply evidence of phonetic and morphemic patterns from nearly 10,000 years ago or so, but it cannot go beyond this period and patterns. Thus, this method also fails to guide the search for the antiquity that could have helped determine the origin of language.

Other speculative linguists imagine that language might have originated from antique systems of communication and information exchange deployed by our ancestors at different junctures of human civilization. To establish their hypothesis, they refer to the similarities observed between animal communication systems and human language. They assume that calls, shouts, and songs of animals might be the sources of human language. If one agrees with this view, one might say that it is true that human language is certainly the result of an evolution of the system of information interchange used in the animal world. In that case, the question of the existence of a "system" in human language is at stake. In the animal world, there are some unique systems of communication, which animals use for fixed biological needs. A honeybee, for instance, when finding a source of pollen, goes back to its hive and dances in a particular manner through which it is able to convey to its fellow bees the kind, direction, and distance of the source. Gibbons, on the other hand, can produce a variety of calls, each being different from the other. Each call has a different connotation, reference, and implication for the members of the clan. This means that a call about the presence of a predator is characteristically different from a mating call or a call made to report the availability of food at some location.

The conclusive inference is that the "language of animals" is fundamentally different from that of humans, because contrary to human language, the "language of animals" is genetically inherited. Moreover, using a fixed number of calls, an animal can share information about different situations or events, such as the position of danger, attack by a predator, location of food, calling for mating, assembling members for collective action, etc. The "language of animals" is a closed world, confined within a fixed set of signals. However, human language is an open-ended world, which is abstract, generative, and free from all kinds of confinement.

We cannot say for sure when and how human language originated or evolved. However, we can definitely say that, when the closed world of signals of the animal world unfolded into an open world of signs and symbols, human language evolved as the most powerful device of thought, expression, and communication.

Difficulties in Language Identification

The issue of language identification within a frame of multilingualism and globalization is not only linked with optical character recognition (OCR) or language technology but also interlinked with more crucial issues relating to language and nation, linguistic identity, ethnology, anthropology, diaspora, and linguistic-cum-cultural imperialism. The rapid growth in literacy, transportation, and communication has significantly increased the number of mother tongues which were once submerged under some dominating languages at various unknown geographical locations but have now surfaced to demand their separate linguistic identities. These languages require the sincere support of modern technology for their survival, growth, and expansion. At present, the world wide web (WWW) is full of texts produced and presented in all major as well as in those so-called minor and unknown languages. In this context, it is necessary to develop a system that can automatically identify a language or variety to attest to and establish its unique linguistic identity in relation to other languages in the global frame. A system is required to identify to which language a particular piece of text document found on the WWW belongs.

There are, however, a variety of issues in tagging information to identify a language as a distinct variety, such as the following:

(a) **Dynamicity:** Because of the dynamic nature of human languages, it is very difficult to obtain complete knowledge of a particular language or variety, and therefore it is nearly impossible to create a static categorization scheme for all natural languages.
(b) **Classification:** This is another difficult issue, as one can use different definitions as working parameters to categorize languages. Also, different purposes may lead one to classify languages in different ways.
(c) **Definition:** Existing systems of language identification are highly inconsistent in their use of the definition of *language*. In many cases, these systems list those features that are not language specific but generic.
(d) **Coverage:** There are, at present, more than 6,800 languages spoken in the world. Present systems of language identification are not powerful enough to cover all languages and scale them properly.
(e) **Documentation:** Existing technology for language identification does not properly document the category to which a particular language or variety belongs. In many cases, the technology provides only the name of a language, which is not adequate for proper identification of a language.

What is understood from the limitations specified above is that any attempt to categorize languages of the world must presume some workable definition of *language*, since there is no single objective definition for *language*. However, when adopting a workable definition to be used in categorizing languages, one has to consider several issues as stated below:

(a) The degree of actual linguistic similarities between speech varieties
(b) Intelligibility among the speakers of the languages in communication with one another
(c) Literacy and ability of the speakers in sharing common literature

(d) Ethnic identities and self-perception of the language communities
(e) Other perceptions and attitudes based on political or social issues

A workable definition may be formulated by combining all the factors stated above. However, the combination of the factors will depend on the needs and purpose of the definition maker.

From the technological point of view, script identification may help in identification of a language or variety. However, this strategy will not work for those languages which have no script or writing system. Moreover, it will lead to wrong identification of languages as there are many languages which use one script; For instance, English, German, French, Spanish, Portuguese, and others use the Roman script, although these are different languages. Furthermore, if the text of a particular language is transcribed into another script, the script-based process of identifying a language can also be deceptive. For these reasons, one has to go beyond the level of script or writing system to trace some more language-specific properties and features that may work in a more reliable manner in the act of language identification. In linguistics, when the script fails to provide the information required for language identification, attempts are made to leverage the direction of writing, vocabulary, language-specific terms and words, idiomatic phrases and expressions, grammar and syntax, place and person names, names of geocultural ideas, concepts, and items, etc., which, due to their unique linguistic identity, provide vital clues for proper identification of a language.

Existing Approaches for Language Identification

As mentioned earlier, in India there are about 22 official languages [70], and 2 or more languages can be written using a single script; For example, the Hindi, Nepali, Rajasthani, Sanskrit, and Marathi languages are written with the Devanagari script, whereas the Bangla, Manipuri, and Assamese languages are written with the Bangla script, etc. Thus, language identification and script identification techniques are different, and the script identification technique cannot be used for language identification, which requires more linguistic information about the languages. The problem of determining the language of a document image has a number of important applications in the field of document analysis, such as indexing and sorting of large collections of such images, or as a precursor to optical character recognition (OCR).

The language identification task can be divided into two categories: (i) image-based language identification, and (ii) text-based language identification. Although many commercial systems bypass language identification by simply looking at the OCR result, it seems better to have a system for language identification which may help to enhance the rate of accuracy of an OCR system. Current methods of discriminating paper documents make use of lexical information derived from optical character recognition (OCR) followed by reference to multiple dictionaries to determine the language. Busch et al. [7] proposed a technique for identification of English, French, and German languages from image documents. The method first

makes generalizations about images of characters, then performs gross classification of the isolated characters and agglomerates these class identities into spatially isolated tokens. Other image-based language identification methods have been proposed by Hochberg et al. [9], Sibun and Spitz [10], Beesley [11], etc.

Although there exist many works on text-based identification of non-Indian languages [11–17], to the best of our knowledge, there has been no effective effort towards text-based identification of Indian languages. There exists only one work on text-based identification for Indian languages of web documents, using the N-gram language model for this purpose [18]. So, there is a need for research in this area for multilingual and multiscript countries such as India.

Script Identification

Script Overview

A script is defined as a graphic representation of the writing system used to write statements and views expressible in a natural language. This implies that a script of a language refers to a particular mode and style of writing and the set of characters used in it. The writing system, which is realized through script, is considered as one of the most versatile strategies used for symbolic representation of speech sounds available in a language. It is designed and deployed in such a manner that it becomes maximally suitable for expressing thoughts, ideas, and views expressible in a language.

A writing system of a natural language is usually endowed with the following six sets of properties [19]:

(a) It contains a set of predefined graphic symbols and signs, which are individually termed as *characters* or *graphemes* and collectively called as *script*.
(b) It contains a well-defined set of rules, which are understood and shared by the members of a speech community for representing the characters in written form.
(c) It includes a set of conventions, which can arbitrarily assign linguistic entity to the symbols, determine their ordering, and define their relations to one another.
(d) It represents the spoken form of a language whose surface constructions and properties are represented, to a large extent, truly by the symbols.
(e) It includes a predefined set of rules which can be recalled for interpreting these symbols for capturing the phonetic entities they represent.
(f) It possesses some physical means for representing these symbols by application to a permanent or semipermanent medium so that all the symbols are interpreted visually as well as through tactile perception.

The study of writing systems of different languages over the centuries has evolved as an independent discipline of investigation (known as *paleography*) in which one tries to investigate and explore the origin of a writing system of a language as well as examine the forms and functions of the individual characters and symbols which are included in the script for writing a language. Within this frame of study, *orthography* demands the full attention of an investigator as it directly relates

to the methods and rules deployed in writing a linguistic form. Since a writing system is a strategy for symbolic representation of thoughts and ideas expressed in speech of a language, it may be considered *complete* to the extent to which it is capable of representing all that may be expressed in the spoken form of the language. A *script*, therefore, is a unique and highly codified collection of characters designed for visual representation of speech sounds used in a natural language. It represents a well-defined and systematic arrangement of distinct graphic symbols (i.e., characters) in specific patterns and orders known as the *alphabet* of a language.

Structural analysis and interpretation of function of the orthographic symbols used in a script of a language are important to the analysis, description, and application of a language because many features of a language are actually encoded in these symbols. In general, information about the form and function of orthographic symbols used in a writing system (or script) of a language becomes indispensable in several areas of applied linguistics, such as *transcription*, *transliteration*, *dictionary compilation*, *language documentation*, *language teaching*, and *language planning* [19]. Also, information in this field becomes useful in several domains of language technology for developing systems and tools for *optical character recognition*, *spoken text transcription*, *cryptography*, *computer keyboard design*, *text-to-speech and speech-to-text conversion*, *linguistic knowledge representation*, *machine learning*, *automatic language recognition*, *language digitization*, etc.

The concepts of *grapheme*, *allograph*, *glyph*, etc. become relevant in the study of the script of a language, because these elements contribute to the formation of the script as well as in designing fonts and typefaces for printing and publication. The term *grapheme* is generally used to refer to the specific atomic units of a given writing system. In this sense, graphemes are *minimally significant* orthographic components which, taken together, form a set of building blocks based on which the text of a given writing system is constructed following the rules of correspondence and usage; For instance, in the English writing system, examples of graphemes include the basic letters of the English alphabet, punctuation marks, and a few other graphic symbols such as those used for numerals and diacritics.

An individual grapheme may be represented in various ways in the script of a language. Each variant may be distinct visually and different in formal features from the others, but all these variants are interpreted as members representing the same grapheme. These individual variants are usually known as *allographs* of the grapheme. It is not mandatory that each language script should have a set of allographs along with the set of graphemes. In fact, there are many language scripts where there are no or only a few allographs; For instance, the Roman script does not have any allographs for the graphemes used in the script. On the other hand, the Devanagari and Bengali scripts possess a set of allographs, which are regularly used as orthographic variants of the graphemes. The preference for an allograph over a grapheme in the act of writing a piece of text is usually controlled by various linguistic and extralinguistic factors, such as the following:

(a) The rules and norms of a writing system of a natural language
(b) The ease and clarity in visual representation of a written text
(c) The goal for minimization of time, labor, and energy in writing

(d) The tools and mediums used for writing a text
(e) The stylistic choices of the text composers
(f) The precision and beauty of the graphic representation of a text
(g) The unconscious features of handwriting of an individual writer

On the other hand, terms such as *glyph* and *sign* are normally used to refer to the physical aspects and properties of a grapheme, allograph, or diacritic symbol used in the script of a language. These are often made up of straight and twisted lines, strokes, curves, hooks, loops, and other kinds of forms, which are combined together in a different order, sequence, and design to form graphemes, allographs, diacritics, and other symbols used in the script.

The evolution of the human thinking process over the millenniums has been instrumental to the origin, development, and evolution of writing systems (i.e., scripts) of languages. The human thinking process has clearly established a strong connection between the linguistic and cognitive abilities of human beings, as people carefully encode and decode information within a written text for successful communication and information interchange. This implies that the script of a language is one of the most effective means for knowledge representation and communication, as orthographic symbols are used to encode and decode knowledge, convert auditory signals into visual images, help to think deductively, transfer information, and order words to construct meaningful sentences.

Since script is defined as a graphic representation of the writing system used to convert auditory statements into visual forms, a *script class*, therefore, refers to a particular style of writing and the order in which the characters are arranged in it. Moreover, as it is already known that there hardly exists a 1:1 mapping between script and language, it is a foregone conclusion that languages across the world are typeset in different scripts. This means that a particular script may be used by many languages with or without any perceivable variation in the shape and size of the alphabets; For instance, the Roman script is used for *English*, *French*, *German*, and some other European languages; the Devanagari script is used for some Aryan languages including *Sanskrit*, *Hindi*, *Bhojpuri*, *Marathi*, etc.; and the Bengali script is used for *Bengali*, *Assamese*, *Manipuri*, and other languages. Moreover, a language can have more than one script for visual representation of its texts; For instance, the *Santhali* language is written in four different scripts, namely the Devanagari, Bengali, Oriya, and Alchiki scripts, while the *Konkani* language is written in the Kannada, Devanagari, and Roman scripts.

Writing systems of languages may differ structurally, genealogically, geographically, stylistically, etc. One can therefore categorize them in several ways, such as according to their type (classifying according to how the system works), family (classifying according to genealogical relations), and region (classifying according to geographical regions). Also, writing systems can be conveniently classified into broad "types" depending on the way they represent their underlying languages. Bearing these issues in mind, writing systems may be divided into six major types [20]: (a) logographic, (b) syllabic, (c) alphabetic, (d) abjads, (e) abugidas, and (f) featural. Details about these types are discussed in ▸Chap. 13 (Middle Eastern Character Recognition) by Abdel Belaid in this handbook.

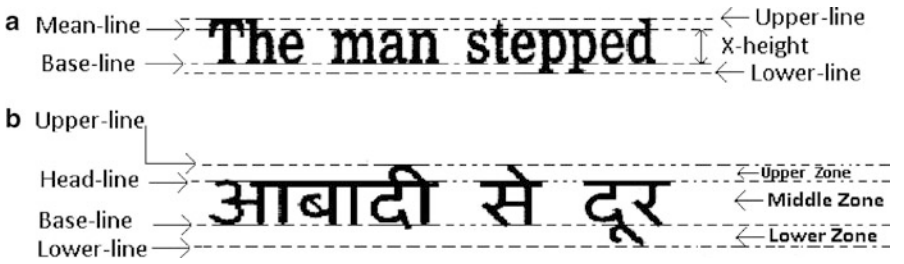

Fig. 9.1 Different zones of (**a**) English and (**b**) Devanagari lines

In the actual act of writing also there are many differences based on the directionality and case type of characters in scripts. While characters in most scripts are written in left to right direction, the characters in Arabic, Urdu, and some other scripts are written from right to left direction. Similarly, in the case of some scripts (e.g., Latin, Cyrillic, etc.) there are both upper-case and lower-case characters, while for some other scripts (e.g., all Indian scripts, Chinese, Japanese, Arabic, etc.) there is no concept of upper and lower case.

One can identify a text line as English upper case if the upper-line and lower-line coincide with the mean-line and base-line, respectively. By the mean-line (or base-line) of a text line we mean the imaginary horizontal line which contains most of the uppermost (or lowermost) points of the characters of a line. We call the uppermost and lowermost boundary lines of a text line as the upper-line and lower-line (Fig. 9.1). From the structural shape of the script characters, we notice that partitioning of Chinese and Arabic text lines into three zones is very difficult, while an English text line can easily be partitioned into three different zones ordered in three tiers: the upper-zone denotes the portion above the mean-line, the middle-zone covers the portion between the mean-line and the base-line, while the lower-zone covers the portion below the base-line. Bengali and Devanagari text lines can also be partitioned into three zones. Different zones in English and Devanagari text lines are shown in Fig. 9.1. Since the shapes of zone-wise features of characters in the Bengali and Devanagari scripts are not similar, these distinguishable shapes are used as the features for proper identification of Bengali and Devanagari script lines.

Single- and Multiscript Documents

Based on its content, a document can be classified as either a single- or multiscript document. If a document contains text of only one script, we call it a single-script document. If a document contains text of two or more scripts, we call it biscript or multiscript. There are many countries where two or more scripts are used; For example, a biscript (Japanese and Roman/English) Japanese document is shown in Fig. 9.2, and a triscript (Bangla, Devanagari, and Roman/English) Indian document

この変形を説明するために Bartlett は，知識を組織化し，記憶中の情報を組織化するための構造である**スキーマ** schemas あるいは**スキマータ** schemata を人間が持っていると仮定した．スキーマや，**フレーム** frames などの知識表現の形式については，後に詳細に解説する．しかし，スキマータもフレームもともに，われわれの記憶がどのように組織化されているかを説明するものである．それらは，語，文，リストを再現する方法についてはあまり考慮しておらず，むしろ記憶のより高度で全体的機能についての側面に焦点を当てている．例えば現実世界の理解にどのようにいたるのか，あるいは，日常的あるいは非日常的な出来事を解釈するために知識をどのように利用しているのか，などがその例である．以下の文章について考えてみよう．

Fig. 9.2 Example biscript document containing English and Japanese text

परवरिश (পর্‌ওঅরিশ্‌, parvarish) *nf*. পালন-পোষণ, লালন-পালন, support, fostering.
परवल (পর্‌ওঅল্‌, parval) *nm*. পটল, a kind of kitchen vegetable.
परवश *see* **परबश** ।
परवा (পর্‌ওআ, parvā) *nf*. প্রত্যেক পক্ষের প্রথম তিথি, the first day of each lunar fortnight.—*nf*. পরোয়া, চিন্তা, care, anxiety.
परवान (পর্‌ওয়ান্‌, parvān) *nm*. প্রমাণ, **सबूत**, proof.
परवानगी (পর্‌ওয়ান্‌গী, parvangī) *nf*. অনুমতি, আদেশ, **इजाज़त**, permission, order.

Fig. 9.3 Example multiscript (English, Bangla, and Devanagari) document

is shown in Fig. 9.3. So, there is a need for the development of multiscript OCR for such countries.

OCR for multiscript document pages can be carried out using one of the following two options: (1) development of a generalized OCR system which can recognize all characters of the alphabets of the possible scripts present in the document pages, or (2) development of a script separation scheme to identify the different scripts present in the document pages with the generation of individual OCR for each script alphabet. Development of a generalized OCR system for multiscript documents is more difficult than single-script OCR development. This is because the features necessary for character recognition depend on the structural

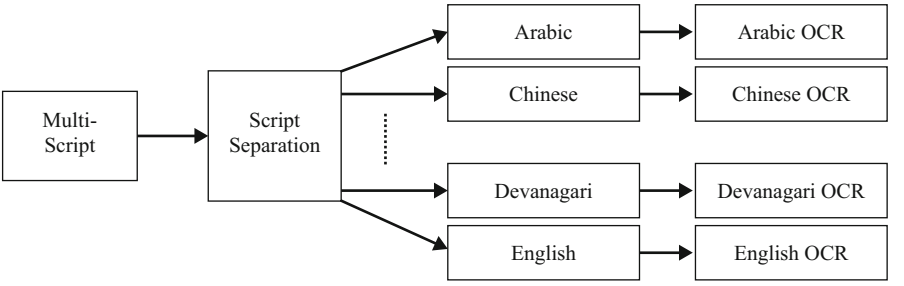

Fig. 9.4 Multiscript OCR technology

properties, style, and nature of writing, which generally differ from one script to another; For example, features used for recognition of Roman/English alphabets are, in general, not good for recognizing Chinese logograms. Also, the large number of characters in all script alphabets is an added difficulty with this technique. On the other hand, the second option is simpler, with multiscript OCR technology generally working in two stages: (1) identification of different script portions from a document, and (2) feeding of individual script portions to the appropriate OCR system. A flow diagram of multiscript OCR is shown in Fig. 9.4.

Script Identification Technology and Challenges

Script identification from a document can be done in different modes such as page-wise, paragraph-wise, line-wise, and word-wise. Word-wise script identification is the most difficult task, as the number of characters in a word is small and hence proper script features from the word may not be obtained for its correct identification. Many pieces of work have been done on script identification from printed text [20], whereas only a few pieces of work have been aimed at identification of different scripts from handwritten documents.

Script identification also serves as an essential precursor for recognizing the language in which a document is written. This is necessary for further processing of the document, such as routing, indexing, or translation. For scripts used by only one language, script identification itself accomplishes language identification. For scripts shared by many languages, script recognition acts as the first level of classification, followed by language identification within the script. Script recognition also helps in text area identification, video indexing and retrieval, and document sorting in digital libraries when dealing with a multiscript environment. Text area detection refers to either segmenting out text blocks from other nontextual regions such as halftones, images, line drawings, etc., in a document image, or extracting text printed against textured backgrounds and/or embedded in images within a document. To do this, the system takes advantage of script-specific distinctive characteristics of text which make it stand out from other nontextual parts of the

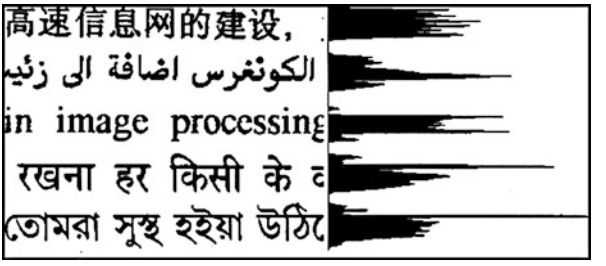

Fig. 9.5 Horizontal projection profiles of a text document containing Chinese, Arabic, English, Devanagari, and Bangla text lines (top to bottom)

document. Text extraction is also required in images and videos for content-based browsing.

One powerful index for image/video retrieval is the text appearing in them. Efficient indexing and retrieval of digital images/videos in an international scenario therefore requires text extraction followed by script identification and then character recognition. Similarly, text found in documents can be used for their annotation, indexing, sorting, and retrieval. Thus, script identification plays an important role in building a digital library containing documents written in different scripts.

Automatic script identification is crucial to meet the growing demand for electronic processing of volumes of documents written in different scripts. This is important for business transactions across Europe and the Orient, and has great significance in a country such as India, which has many official state languages and scripts. Due to this, there has been growing interest in multiscript OCR technology during recent years.

Script identification relies on the fact that each script has a unique spatial distribution and visual attributes that make it possible to distinguish it from other scripts. So, the basic task involved in script recognition is to devise a technique to discover such distinctive features from a given document and then classify the document's script accordingly.

Many features have been proposed for script identification. Some of the features used in script identification are discussed below. Projection profile is one of the important features used by various researchers for line-wise script identification; For example, see Fig. 9.5, where the horizontal projection profiles of a text document containing Chinese, Arabic, English, Devanagari, and Bangla text lines are shown. From this figure it can be seen that some of the text lines have different behavior in their projection profile. Arabic, Devanagari, and Bangla text lines have only one, large peak in their profile, while Chinese and English do not have this feature but rather have two or more peaks of similar height. Also it can be noted that, in Arabic, the peak occurs at the middle of the text line, while in Bangla and Devanagari the peak occurs at the upper half of the text line.

Run information is also used as a feature in script identification. The maximum run for each row can provide a distinct feature to classify it from others; see Fig. 9.6, where the row-wise longest horizontal run is shown. It can be seen from this figure that Bangla and Devanagari text has a very large run compared with others, which is due to touching of characters in a word through the head-line. Because of the

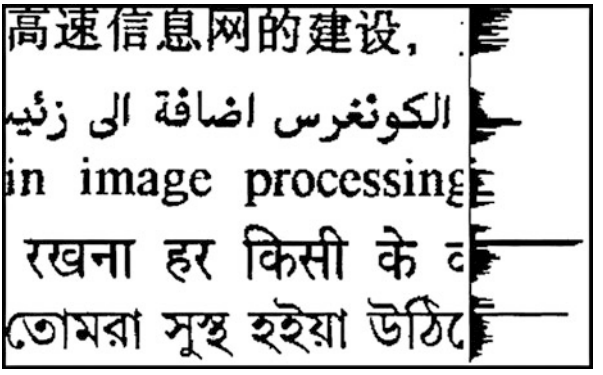

Fig. 9.6 Row-wise longest horizontal run in Chinese, Arabic, English, Devanagari, and Bangla text lines (from top to bottom)

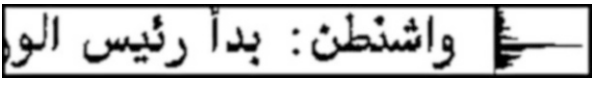

Fig. 9.7 Example of an Arabic text line where a long horizontal run is obtained. Note that, here, the long horizontal run appears at the lower-half of the text line

base-line, sometimes large runs may be obtained from Arabic lines (as seen in Fig. 9.7), but its position is different from others. In Arabic, large runs generally occur in the base-line part, while in Bangla and Devanagari they occur in the mean-line portion.

The crossing count is another important feature for identifying some scripts from others; For example, Chinese script has a higher crossing count than English or Arabic. However, the crossing feature is not very robust due to noise where images may be broken or touched. To tackle this problem, a run length smoothing algorithm (RLSA) concept has been used for script line identification. The RLSA is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's. The algorithm transforms a binary sequence X into an output sequence Y according to the following rules: (1) 0's between two consecutive 1's in X are changed to 1's in Y if the number of 0's is less than or equal to a predefined limit L. (2) 1's in X are unchanged in Y. (3) 0's at the boundaries of X are unchanged in Y. For example, in row-wise RLSA with $L = 5$, the sequence X is mapped into Y as follows:

$$X : 0000110000000100100010010000000011100$$

$$Y : 0000110000000111111111110000000011100$$

The RLSA is applied row by row as well as column by column to a text line, yielding two distinct bit maps which are then combined by a logical AND operation. The original text lines and the result of applying RLSA are shown in Fig. 9.8. The parameter L can be fixed using text line height information.

Fig. 9.8 Original text lines (*top*) and modified RLSA results (*bottom*) for (**a**) Chinese, (**b**) Roman, and (**c**) Arabic script

From Fig. 9.8 it can be noted that most of the character portion gets filled up when the RLSA is applied to a Chinese text line, which is not true for English or Arabic text lines. The percentage of character filled-up area of the RLSA output text line with respect to its original version can be used as a measure for script identification. Based on this feature, a Chinese text line can easily be separated from an English one.

Features based on the water reservoir principle are also widely used in script identification. The water reservoir principle is as follows: If water is poured from a given side of a component, the cavity regions of the component where water will be stored are considered as the reservoirs of that component. Characters in a word may touch, and thus two touching consecutive characters in a word create large cavity regions (space) between them and hence large reservoirs [21]. Details of the water reservoir principle and its different properties can be found in [21]. However, here we give a short description of different properties used in the scheme to ease readability.

Top (bottom) reservoir: By the top (bottom) reservoir of a component we mean the reservoir obtained when water is poured from the top (bottom) of the component. The bottom reservoir of a component is visualized as a top reservoir when water is poured from the top after rotating the component by 180°.

Left (right) reservoir: If water is poured from the left (right) side of a component, the cavity regions of the component where water will be stored are considered as the left (right) reservoir. The left (right) reservoir of a component is visualized as a top reservoir when water is poured from the top after rotating the component by 90° clockwise (anticlockwise).

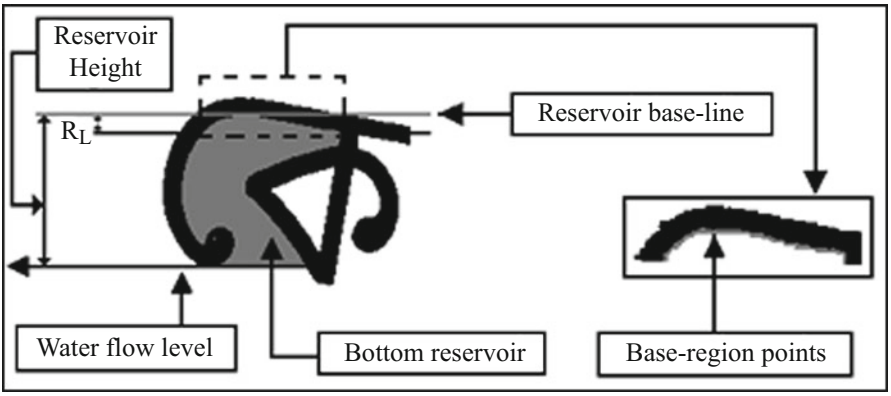

Fig. 9.9 Reservoir base-line and water flow level in a bottom reservoir. Base-region points of the bottom reservoir are marked *grey* in the zoomed part of the component

Water reservoir area: By the area of a reservoir we mean the area of the cavity region where water will be stored. The number of points (pixels) inside a reservoir is computed, and this number is considered as the area of the reservoir.

Water flow level: The level from which water overflows from a reservoir is called the water flow level of the reservoir (Fig. 9.9).

Reservoir base-line: A line passing through the deepest point of a reservoir and parallel to the water flow level of the reservoir is called the reservoir base-line (Fig. 9.9).

Height of a reservoir: By the height of a reservoir we mean the depth of water in the reservoir. In other words, the height of a reservoir is the normal distance between the reservoir base-line and the water flow level of the reservoir (Fig. 9.9).

Base-region points: By the *base-region* points of a reservoir we mean some of the reservoir's surface (border) points whose perpendicular distances are less than R_L from the reservoir base-line. Here, R_L is the stroke width of the component. *Base-region* points of a bottom reservoir are marked by grey in Fig. 9.9. The stroke width (R_L) is the statistical mode of the black run lengths of the component. For a component, R_L is calculated as follows: A component is first scanned row-wise (horizontally) and then column-wise (vertically). If n different runs of length $r_1, r_2, \ldots r_n$ with frequencies $f_1, f_2, \ldots f_n$, respectively, are obtained after these two scans of the component, then the value of R_L will be r_i if $f_i = \max(f_j)$, $j = 1, 2, \ldots n$.

Different reservoir-based features can give various distinctive properties that are useful for script identification. Some of these distinctive features are discussed below.

From Fig. 9.10, it can be seen that the water flow level of the top reservoirs of most English characters is the top of the character, whereas this is not true for the characters of the Arabic script. This distinctive feature obtained based on the reservoir is very helpful for identification of Roman and Arabic.

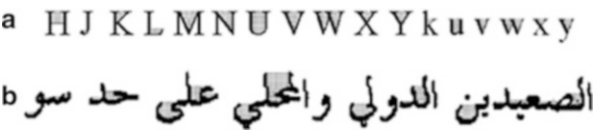

Fig. 9.10 Example water reservoirs obtained for some (**a**) English and (**b**) Arabic characters. Here, reservoirs are marked by *dots*

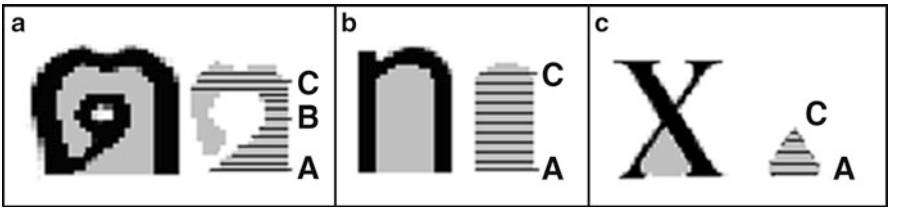

Fig. 9.11 Shape of bottom reservoir shown in (**a**) Thai and (**b**, **c**) English characters. Reservoir widths of different rows are shown by *horizontal lines*

Bottom reservoirs also play an important role in script identification; For example, there is a distinct difference between the shapes of the bottom reservoirs in some English and Thai characters, and this reservoir-based shape difference helps in identification of Roman and Thai scripts. To find this shape difference, we compute the width of the reservoir in every row and note the change of these widths. (By the reservoir width of a row we mean the distance between two border points of that reservoir in that row.) Because of the irregular shape of the water reservoir of Thai characters, the reservoir width for different rows will be different as we move towards the bottom of the reservoir starting from the water flow level. However, for English characters, the reservoir width for different rows will either remain the same or tend to decrease as we move upwards from the reservoir flow level, because of the regular shape of the reservoir for English characters. For an illustration, see Fig. 9.11, which shows the shape of the bottom reservoir of one Thai and one English character. For the Thai character, the reservoir width decreases from A to B and again increases from B to C. For the English character, the reservoir width generally remains the same or decreases from A to C.

Run number information is also used as a feature for script identification; For example, one of the most distinctive and inherent characteristics of the Chinese script is the existence of many characters with four or more vertical black runs. In the English and Arabic alphabets, the number of such characters is very low. If the character-wise maximum run numbers of English, Arabic, and Chinese text lines are computed, such distinctive behavior can indeed be observed; For example, see Fig. 9.12, where run statistics are computed from 50,000, 40,000, and 35,000 characters of English, Chinese, and Arabic text, respectively. From this figure it can be observed that the graphs for Arabic and English show similar behavior after run number 3. Also, it can be observed that the graph obtained from the Chinese text has a different nature. In a Chinese text line, about 57 % of characters have four or more black runs. In English, there is only one character (“g”) which has four vertical runs.

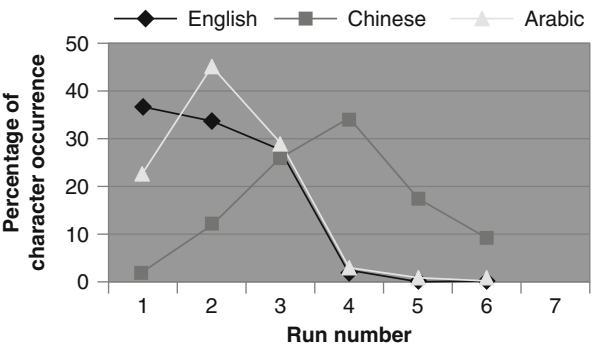

Fig. 9.12 Run numbers with their occurrence percentage of characters for English, Chinese, and Arabic texts

We observe that the percentage of characters with one or two vertical black runs in a Chinese text line is very low (about 14 %), whereas in English and Arabic text there are many characters (more than 65 %) with one or two runs. In this regard, Chinese script shows different behavior from the English and Arabic scripts. This simple but important criterion can be used for separation of Chinese from English and Arabic texts.

Among other important features, moments, the concavity distribution, textural features, fractal features, GLCM, wavelets, Gabor features, etc. are used [20]. Different classifiers such as LDA, MLP, KNN, GMM, SVM, etc. are used for different script identification purposes.

Machine-Printed Script Identification

Among the pieces of earlier work on separation of different scripts, the first attempt to address this problem was due to Spitz [22–24]. In 1990, he described a method for classification of individual text from a document containing English and Japanese script [22]. Later, he [24] proposed a technique for distinguishing Han- and Latin-based scripts on the basis of the spatial relationships of features related to the character structures. Identification within the Han-based script class (Chinese, Japanese, Korean) is performed by analysis of the distribution of optical density in text images. Latin-based script is identified using a technique based on character shape codes, and finding the frequency of language-specific patterns.

Different scripts have distinctive visual appearance. So, a block of text may be regarded as a distinct texture pattern, and this property has been used for script identification. Generally, texture-based approaches are used for block-wise script identification and do not require component analysis. Thus, a texture-based approach may be called a global approach. Representative features for each script are obtained by average values of texture features extracted from training document images. For an input document written in unknown script, similar features are extracted. These features are then compared with the representative features of each script in order to identify the script of the input document. Many researchers have

used texture features for script identification [7, 25–27]. An identification scheme using cluster-based templates for scripts was proposed by Hochberg et al. [25, 26]. They developed a script identification scheme for machine-printed documents in 13 scripts: Arabic, Armenian, Burmese, Chinese, Cyrillic, Devanagari, Ethiopic, Greek, Hebrew, Japanese, Korean, Latin, and Thai. Here, templates for each script were created by clustering textual symbols from a training set. Script identification was then done by comparing textual symbols of an unknown document with those templates. Busch et al. [7] evaluated different textural features such as gray-level co-occurrence matrix features, wavelet energy features, wavelet log-mean deviation features, wavelet co-occurrence signatures, wavelet scale co-occurrence signatures, etc. for the purpose of script identification. They considered English, Chinese, Greek, Cyrillic, Hebrew, Hindi, Japanese, and Persian scripts for their experiment.

Gabor filtering has been commonly used for script identification from printed documents [27, 28]. Using rotation-invariant texture features, Tan [27] described a method for identification of Chinese, English, Greek, Russian, Malayalam, and Persian text. Rotation-invariant texture features are computed based on an extension of the popular multichannel Gabor filtering technique, and their effectiveness was tested with 300 randomly rotated samples of 15 Brodatz textures. These features were then used for script identification from machine-printed documents. Recently, Pam et al. [29] proposed a technique for identification of Chinese, Japanese, Korean, and English scripts using steerable Gabor filters. Here, a Gabor filter bank is first appropriately designed so that rotation-invariant features can handle scripts that are similar in shape. Next, the steerable property of Gabor filters is exploited to reduce the high computational cost resulting from the frequent image filtering, which is a common problem encountered in Gabor-filter-related applications. Gabor filters have also been used to identify scripts from handwritten documents [9].

Ding et al. [30] proposed a method for separating two classes of scripts: European (comprising Roman and Cyrillic scripts) and Oriental (comprising Chinese, Japanese, and Korean scripts), using several discriminating structural and statistical features. Zhang and Ding [31] presented a cluster-based bilingual (English and Chinese) script identification approach. The approach proposed by Lee et al. [32] focuses heavily on reliable character extraction and then uses a variety of decision procedures, some handcrafted and others trainable, to detect document scripts. Ablavsky and Stevens [33] proposed an approach for Russian and English documents that applies a large pool of image features to a small training sample and uses subset feature selection techniques to automatically select a subset with the greatest discriminating power. At run time, they used a classifier coupled with an evidence accumulation engine to report a script label, once a preset likelihood threshold had been reached. Elgammal and Ismail [34] proposed a scheme for identification of Arabic and English text from a document. Features obtained mainly from horizontal projection profiles and run length histograms are used in their scheme. Wood et al. [35] described an approach using filtered pixel projection profiles.

Fractal-based features are also used for script identification. Tao and Tang [36] proposed an approach based on the modified fractal signature (MFS) and modified fractal features (MFF) for discrimination of Oriental and Euramerican scripts.

India has 22 official languages, and 11 scripts are used. In India many documents are written in two or more scripts/languages. Many features, e.g., Gabor features, texture features, water reservoir-based features, head-line, run length smoothing algorithm, edge-based features, directional features, convexity, etc. have been used. To provide an outlook for the reader, details of work done in the Indian scenario are presented below.

The existing pieces of work on Indian script identification mainly rely on local approaches and can be divided into two types. One type of work deals with line-wise script identification. Here, it is assumed that a single line contains only one script. Another type of work deals with word-wise script identification. Here, it is assumed that a single text line may contain words in two or more scripts. In this case, each text line is segmented into individual words and the script of each word of a line is then identified. Word-wise script identification is more difficult and challenging than line-wise identification because of the smaller number of characters and hence lower amount of information obtained.

Although India is a multilingual and multiscript country, generally a document contains three languages. Under the three-language formula, the document may be printed in English, Devanagari, and one of the other Indian state official languages; For example, a document of the West Bengal State government contains English, Devanagari, and Bangla text. To take care of such trilingual (triplet) documents, Pal and Chaudhuri [37] proposed an automatic scheme for separating text lines for almost all triplets of script forms. To do so, the triplets are grouped into five classes according to their characteristics, and shape-based features are employed to separate them without any expensive OCR-like algorithms. Later, they [38] proposed a generalized scheme for line-wise identification of all Indian scripts from a single document.

Chaudhuri and Sheth [39] proposed a trainable script identification strategy for Indian documents containing Devanagari, English, and Telugu scripts. In this work, three trainable classification schemes are proposed for identification of Indian scripts. The first scheme is based upon a frequency-domain representation of the horizontal profile of the textual blocks. The other two schemes are based on connected components extracted from the textual region. These schemes exploit properties of the Gabor filter and the width-to-height ratio of the connected components.

Pal and Chaudhuri [40] proposed an automatic technique for identification of different script lines from a printed document containing the five most popular scripts in the world, namely Roman, Chinese, Arabic, Devanagari, and Bangla. Here, the head-line information is first used to separate Devanagari and Bangla script lines into one group and English, Chinese, and Arabic script lines into another group. Next, from the first group, Bangla script lines are distinguished from Devanagari using some structural features. Similarly, using the vertical run number and run length smoothing [41], the Chinese text lines are identified from the second

group, and then using features obtained from the concept of the water reservoir overflow analogy [42] as well as statistical features, English text lines are separated from Arabic.

One of the pioneering works on word-wise script identification for Indian languages may be credited to Pal and Chaudhuri [43]. Here, *Shirorekha* features and other different stroke features such as vertical line, horizontal line, slanted line, angular shapes, etc. are used in a tree classifier for word-wise identification of Devanagari, English, and Bangla scripts.

Patil and Subbareddy [44] proposed a neural network-based system for word-wise identification of English, Hindi, and Kannada language scripts. In this system, features are first extracted in two stages. In the first stage, the document image is dilated using 3×3 masks in horizontal, vertical, right diagonal, and left diagonal directions. In the second stage, the average pixel distribution is found in these resulting images. Based on these features, a modular neural network technique is used for the identification.

Dhanya et al. [45, 46] proposed a Gabor filter-based technique for word-wise segmentation of bilingual documents containing English and Tamil scripts. They proposed two different approaches. In the first method, words are divided into three distinct spatial zones. The spatial spread of a word in the upper and lower zones, together with the character density, is used to identify the script. The second method analyzes the directional energy distribution of a word using Gabor filters with suitable frequencies and orientations.

Manthalkar and Biswas [47] proposed a rotation-invariant texture classification technique for script identification from Indian documents containing Bangla, Kannada, Malayalam, Oriya, Telugu, Marathi, and English.

Handwritten Script Identification

Although there are many pieces of work on script identification from printed documents, only a few pieces of work are available for handwritten documents.

Hochberg et al. [9] proposed a feature-based approach for script identification in handwritten documents and achieved 88 % accuracy from a corpus of 496 handwritten documents in distinguishing Arabic, Chinese, Cyrillic, Devanagari, Japanese, and Latin. In their method, a handwritten document is characterized in terms of the mean, standard deviation, relative vertical centroid, relative horizontal centroid, number of holes, sphericity, and aspect ratio of the connected components in a document page. A set of Fisher linear discriminants (FLDs), one for each pair of script classes, is used for classification. The document is finally assigned to the script class to which it is classified most often.

Fractal features have been used by Roy et al. [49] for script separation of handwritten documents. A fractal [49] is defined as a set for which the Hausdorff–Besikovich dimension is strictly larger than the topological dimension. The fractal dimension is a useful method to quantify the complexity of feature details present in an image. The fractal dimension is an important characteristic of fractals because

it contains information about their geometric structures. When employing fractal analysis, researchers typically estimate the fractal dimension from an image and use it for the desired purpose.

Although Gabor function-based script recognition schemes have shown good performance, their application is limited to machine-printed documents only. Variations in writing style, character size, and interline and interword spacing make the recognition process difficult and unreliable when these techniques are applied directly to handwritten documents. Therefore, it is necessary to preprocess document images prior to application of the Gabor filter to compensate for the different variations present. This has been addressed in the texture-based script identification scheme proposed in [49]. In the preprocessing stage, the algorithm employs denoising, thinning, pruning, m-connectivity, and text size normalization in sequence. Texture features are then extracted using a multichannel Gabor filter. Finally, different scripts are classified using fuzzy classification. In this proposed system, an overall accuracy of 91.6 % was achieved in classifying documents handwritten in four different scripts, namely Latin, Devanagari, Bengali, and Telugu.

Finally, to give an idea of the recognition accuracies of different systems, Table 9.1 summarizes some of the benchmark work reported for printed and handwritten script recognition. Various features and classifiers along with the scripts used by different researchers are also listed in the table, as mentioned in [20].

Font and Style Recognition

A language, if it has a writing system, invariably has a set of typographic symbols and signs (called characters) to represent the sounds used in the language. The complete set of these symbols is known as script, in which each character is rendered in different typefaces, shapes, sizes, designs, and angles – collectively known as font. This means that a font may be defined as a combination of sorts used to compose characters with a unique design, size, and style for a particular script. This signifies that a language can have several fonts to represent the same script; For instance, the Roman script has several fonts developed in different shapes, sizes, and designs, although all of them collectively represent the same script. Moreover, a font of a particular design may vary from its nearest peer based on certain characteristically discernible properties such as height,[1] weight,[2] width,[3] size,[4] slope,[5] curvature, thickness, boldness, etc., which are treated as feature-based parameters in identification of a particular font from the pool of many fonts used for a script.

The differences in fonts are necessitated due to various reasons – starting from personal choice to text content to text readability to typeface compatibility with a computer system. The font used in legal texts, for instance, may vary in design and shape from the font used in medical texts or in texts meant for use by young language learners. Such variation in font use for different text types triggers the question of choice of font and style of writing and also posits problems for font and style recognition by both man and machine, as proper recognition of a font often contributes to correct recognition of a text vis-à-vis a language.

Table 9.1 Script identification results using different methods applied to printed and handwritten documents

Method	Feature	Classifier	Script type	Document type	Script identification mode	Identification accuracy
Splitz [24]	Upward concavity distribution	Var. comparison	Latin, Han	Printed	Page-wise	100 %
	Optical density	LDA + Eucl. dist.	Chinese, Japanese, Korean			
Lam et al. [50]	Hor. proj., height distribution	Stat. classifier	Latin, Oriental scripts	Printed	Page-wise	95 %
	Circles, ellipses, ver. strokes	Freq. of occurr.	Chinese, Japanese, Korean			
Hochberg et al. [25]	Textual symbols	Hamming dist. classifier	Arabic, Armenian, Devanagari, Chinese, Cyrillic, Burmese, Ethiopic, Japanese, Hebrew, Greek, Korean, Latin, Thai	Printed	Page-wise	96 %
Hochberg et al. [51]	Textual symbols	Hamming dist. classifier		Printed	Word-wise	Not available
Pal et al. [52]	Headlines, strokes, ver. runs, lowermost pt., water resv.	Freq. of occurr.	Devanagari, Bengali, Chinese, Arabic, Latin	Printed	Line-wise	97.33 %
Elgammal et al. [34]	Hor. proj. peak, moments, run-length distribution	Feedforward NN	Arabic, Latin	Printed	Line-wise	96.80 %

Jawahar et al. [54]	Headline, context info.	PCA + SVM	Devanagari, Telugu	Printed	Word-wise	92.3–99.86 %
Chanda et al. [55]	Headline, ver. strokes, tick left/right profiles, water resv., deviation, loop, left incline	Freq. of occurr.	Devanagari, Bengali, Latin, Malayalam, Gujurati, Telugu	Printed	Word-wise	97.92 %
Wood et al. [35]	Horizontal/vertical proj.	–	Arabic, Cyrillic, Korean, Latin	Printed	Page-wise	Not available
Jain et al. [56]	Texture feature using discriminating masks	MLP	Latin, Chinese	Printed	Page-wise	Not available
Tan [27]	Gabor filter-based texture feature	Weighted Eucl. dist.	Chinese, Greek, Malayalam, Latin, Russian, Persian	Printed	Page-wise	96.70 %
Peake et al. [28]	GLCM features	KNN classifier	Chinese, Greek, Malayalam, Latin, Russian, Persian, Korean	Printed	Page-wise	77.14 %
	Gabor filter					95.71 %
Singhal et al. [57]	Gabor filter-based texture feature	Fuzzy classifier	Devanagari, Bengali, Telugu, Latin	Handwritten	Page-wise	91.60 %

(*continued*)

Table 9.1 (continued)

Method	Feature	Classifier	Script type	Document type	Script identification mode	Identification accuracy
Busch et al. [7]	GLCM features	LDA + GMM	Latin, Chinese, Japanese, Cyrillic, Greek, Devanagari, Hebrew, Persian	Printed	Para-wise	90.90 %
	Gabor energy					95.10 %
	Wavelet energy					95.40 %
	Wavelet log mean dev.					94.80 %
	Wavelet co-occurrence					98 %
	Wavelet log co-occurrence					99 %
	Wavelet scale co-occurrence					96.80 %
Joshi et al. [58]	Gabor energy distribution, horizontal projection profile, energy ratios	KNN classifier	Devanagari, Latin, Gurmukhi, Kannada, Malayalam, Urdu, Tamil, Gujurati, Oriya, Bengali	Printed	Para-wise	97.11 %
Ma et al. [59]	Gabor filter-based texture feature	KNN + SVM multiclassifier	Latin, Devanagari	Printed	Word-wise	98.08 %
			Latin, Arabic			92.66 %
Dhanya et al. [46]	Gabor filter-based directional feature	SVM	Tamil, Latin	Printed	Word-wise	96 %

Jaeger et al. [60]	Gabor feature	SVM	Arabic, Roman	Printed	Word-wise	90.93 %
		SVM	Chinese, Roman			93.43 %
		KNN	Korean, Roman			94.04 %
		KNN	Hindi, Roman			97.51 %
Chanda et al. [61]	Structural feature based on background and foreground information	SVM	Roman, Thai	Printed	Word-wise	99.62 %
Zhou et al. [62]	Connected component-based feature	Rule-based classifier	Bangla, Roman	Printed	Block-wise (postal address block)	99.00 %
		Rule-based classifier	Bangla, Roman	Handwritten		95.00 %
Dhandra et al. [63]	Density feature after morphological operation	Nearest neighbor	Roman, Hindi, Kannada, Urdu	Printed	Page-wise	97.00 %
				Handwritten		85.46 %
Hochberg et al. [9]	Hor./ver. centroids, aspect ratio, white holes	FLD	Arabic, Chinese, Cyrillic, Devanagari, Japanese, Latin	Handwritten	Page-wise	88 %

Keeping these crucial issues in view, at attempt is made in this section to refer to some of the methods and approaches developed for font and style recognition in written texts. In the following subsections, attention is concentrated on defining the term "font" within a general conceptual frame, methods of font generation, issues in variation of font formation, and recognition strategies for font and style.

Font Terminology

In typography, a font is a set of printable or displayable text characters of a single size and style of a particular typeface [64]. The overall design of the character shapes is determined by the typeface, and the style refers to the average stroke width of characters [boldface versus lightface (normal)] and the posture of the body (italic versus Roman). Points or picas are used as units for font size. The size of a font is typically given in points (1 in. = 72 points, 1 in. = 6 picas). Pitch information is also used for size, where pitch means the number of characters displayed per inch. The set of all the characters for 9-point Times New Roman italics is an example of a font. Similarly, the set of all the characters for 10-point Times New Roman italics should be treated as a separate font, as the size is different here.

Font Generation

In this age of computer technology, it is not a difficult task to design and develop a large number of fonts in digital form for many of the language scripts. According to [64], there are two basic ways to generate a font with a computer, as follows: In the first method, font artwork is scanned and quantized into bitmaps by a scanner, and then the font contours, represented by spline knots [64], are traced out from the high-resolution bitmaps and stored on a disk. The spline knot representation not only is an effective way of achieving data compression but also can serve as an effective master for later processing. In the second method, without using an artwork master, the spline knots are marked on a display screen and then entered with the desired stroke width. Construction of spline curves that pass through given spatial points to form the median or skeleton of the desired font bitmaps can be carried out using this procedure. In bitmap reconstruction, an imaginary electronic paintbrush with elliptical cross-section is swept along the spline curves, rotating along the way with angles complying with the stroke-width information. In this way, a font whose strokes are represented by the spline functions may be generated completely independently without any artwork [64].

Font Variation

In the present computer age, the wide variety of fonts available may be organized into groups called "font families." The fonts within a font family show typeface

similarities. As discussed in [75], the members of a font family may differ from each other in that one of them may be bold, another italic or use small caps, etc.; For example, the original Arial font, the Arial Black font, which is bold, and Arial Narrow, which is thinner than the rest, all come under the Arial font family. Some of the most popular font families include Times New Roman, Helvetica, Arial, etc. A font is fully specified by five attributes as noted in Hou [64]: typeface (Times, Courier, Helvetica), weight (light, regular, demibold, heavy), slope (Roman, italic), width (normal, expanded, condensed), and size. See Fig. 9.13 for examples of some fonts.

A very popular classification is that based on five generic font families, which is used for the purposes of cascading style sheets (CSSs) – a widely used mechanism for adding style (e.g., fonts, colors, spacing, etc.) to web documents [75]. The five generic font families consist of serif fonts (Times New Roman, MS Georgia, etc.), sans-serif fonts (MS Trebuchet, Univers, etc.), cursive fonts (Adobe Poetica, Sanvito, etc.), decorative fantasy fonts (Critter, Cottonwood, etc.), and fixed-width monotype fonts (Courier, MS Courier New, etc.).

Recognition Strategies for Font and Style

Optical font recognition (OFR) is an important issue in the area of document analysis. It aims at recovering typographical attributes with which texts have been printed. Although this is a difficult and time-consuming task, it is an important factor for both character recognition and script identification for multilingual text documents. The number of alternative shapes for each class can be reduced by font classification, leading to essentially single-font character recognition [65]. This helps to improve OCR accuracy. To achieve automatic typesetting, the output of a document processing system should include not only the content of the document but also the font used to print it, and font recognition helps to do this.

Due to its usefulness for both document and character recognition, OFR can be considered to be a challenging problem to reduce the search space of word-to-word matching in keyword spotting. Besides the improvement in OCR and document indexing and retrieval, OFR has many other valuable applications; For example, font identification could be used as a preprocessing step in an automated questioned document analysis process. For crime detection, a forensic expert might be interested in analyzing only documents that are typeset with a particular font. The preprocessing step will help the forensic expert to narrow down the search space to find relevant evidence. Moreover, identifying the font of a document might help an investigator to determine its source of origin (since fonts are sometime country specific).

Font is also used in recognition of logical document structures. Here, knowledge of the font in a word, line, or text block may be useful for defining its logical label, such as chapter title, section title, paragraph, etc. Font information is also required for document reproduction, where knowledge of the font is primarily used in order to reprint a similar document.

A	Wide latin	A	Harrington
A	Vladimir	A	Harlow Solid Italics
A	Vivaldi	A	Gigi
A	Viner Hand ITC	A	Freestyle Script
A	STENCIL	A	Arial Narrow
A	Snap ITC	A	Blackadder ITC
A	Ravie	A	Bradley Hand ITC
A	Pristina	A	Broadway
A	Times new Roman	A	Brush Script MT
A	Papyrus	A	Century Gothic
A	Old English Text MT	A	Chiller
A	Monotype Corsiva	A	Colonna MT
A	Mature MT Script Capitals	A	Copper Black
A	Magneto	A	Courier
A	Lucida Fax	A	Edwardin Script ITC
A	Lucida Caligraphy	A	ALGERIAN
A	Kunstler Script	A	Coronet
A	Kristen ITC	A	EUCLID MATH TWO
A	Jokerman	A	Helvetica Narrow
A	Informal Roman	A	Comic Sans MS

Fig. 9.13 Some examples of different fonts

There are mainly two possible approaches for font recognition: global and local [67]. Global features are generally detected by nonexperts in typography (text density, size, orientation and spacing of letters, serifs, etc.). This approach is suitable for *a priori font recognition*, where the font is recognized without any knowledge of the letter classes. On the other hand, in the local approach, features are extracted

from individual letters. The features are based on particularities of letters such as the shapes of serifs (coved, squared, triangular, etc.) and the representation of particular letters at different angles, e.g., "g" and "*g*," and "a" and "*a*." This kind of approach may derive substantial benefit from knowledge of the letter classes.

In spite of the many applications of automatic font recognition, few pieces of work have addressed these issues. As mentioned earlier, local as well as global features are used by researchers for font recognition. Typological features such as font weight (reflected by the density of black pixels in a text line image), slopes, sizes, typefaces, etc. [66], texture features [68], connected component-based features, and clusters of word images [69] are used for font recognition. Most of the available work on font recognition has been done on English; using 32 English fonts, 99.1 % accuracy was obtained by Khoubyari and Hull [68]. Some pieces of work have been done for Arabic, Chinese, and Japanese [72]. Although pieces of work exist on non-Indian languages, work on Indian script font detection is very limited in spite of the fact that many scripts and languages are used in India [76]. Some pieces of work have been done towards detection of different styles such as bold, italics, etc. For boldface detection, density features are used, whereas for italics detection, slant estimation techniques are incorporated.

Morris [72] proposed a study on the problem of digital font recognition by analyzing Fourier amplitude spectra extracted from word images. This study was mainly performed in order to examine the applicability of human vision models to typeface discrimination and to investigate whether spectral features might be useful in typeface production. First, Fourier transformation is applied to the word image, then a vector of features is extracted by applying many filters to the resulting spectra. A quadratic Bayesian classifier is used for font classification and showed good results considering the many important simplifications. Chanda et al. [76] defined models for characters and placed them in a tree according to certain attributes (ascenders, descenders, holes, etc.). Some preprocessing was done on the text in order to select one sample for each kind of shape (letter). The selected shapes were placed in the tree according to their attributes. In another operation, a tree was generated for each font in the system (instantiation of a generic tree with the different letters of the given font). Next, the shape tree was compared with each font tree in order to compute a distance in order to obtain the associated font. The objective of the font identification was mainly to improve the performance of the OCR system by limiting the search space.

Nowadays, some commercial software is available for font recognition; CVISION (http://www.cvisiontech.com/reference/ocr/font-recognition-software.html?lang=eng) is one such piece of software.

Conclusions

This chapter deals with issues of language identification, script identification, and font-cum-style recognition – three important domains in the document analysis area. These domains are so closely interlinked with each other that discussion on one domain without reference to the other two domains is bound to be skewed

and incomplete. Therefore, all three domains were adequately addressed within the frame of a single chapter. After an overview about language in general, the chapter discussed, in a step-by-step process, the origin of language, problems faced in language identification, and approaches used in language identification. The second section presented an overview of script, showed differences between single- and multiscript documents, discussed issues and challenges involved in script identification, defined strategies for printed script identification by machines, and focused on methods for handwritten script identification. The final section defined terminologies used in automatic font recognition, discussed problems of font generation, presented font variations, and highlighted font and style recognition strategies.

In essence, this chapter presents an insightful survey on the problems and issues relating to the development of technology for machine recognition of language, script, and font of text documents. It also provides glimpses on the present state of the art of the technology to draw the attention of readers towards new directions.

In the context of script identification, it may be mentioned that most works on non-Indian languages to date have been based on either line-wise or paragraph/block-wise methods (a block being nothing but a rectangular box containing portions of several consecutive text lines of the same script). This technology, however, does not exert much impact in the context of Indian languages, where multiscript text documents are the rule of the day. Since the lines of text documents in many of the Indian languages contain two or more language scripts, it becomes necessary to design a word-based script identification technique that will yield a higher level of accuracy in case of multiscript documents. Moreover, there is an urgent need for rigorous operation on poor and noisy images of Indian text documents, since most existing script identification techniques operating on Indian languages have considered only clear and noiseless text documents.

On the other hand, there are many handwritten text documents where there is a need for application of script identification techniques. A postal document produced in India, for instance, may be written by hand in any of the Indian languages. In this case, we need to design an automatic postal sorting machine to recognize pin codes and city names. There exists only one work on word-based identification of handwritten Indian documents containing Bangla and Devanagari scripts [49]. This technique may be customized for other Indian and foreign-language scripts, with wider application potential across the globe.

Finally, we assume that online script identification techniques may become highly useful with extensive use of pen-computing technology. However, since these techniques have not received much attention to date, more work is needed in this direction. A similar observation is valid for automatic script identification technology from video-based text documents.

Cross-References

▸Asian Character Recognition
▸Middle Eastern Character Recognition

Notes and Comments

[1]Height refers to the vertical length of a character between the two imaginary lines drawn on its head and foot, respectively.

[2]Weight signifies the thickness of the character outlines relative to its height.

[3]Width refers to the horizontal stretch of a character within the specified space allotted to it for its formation and visual representation.

[4]Size refers to the coverage of space of a character in accordance to its degree of readability.

[5]Slope is connected with the angle at which a particular character is made to stand in a text made in linear order. It can be upright, straight or slanted in right or left direction.

References

1. Greenberg J (1963) Universals of languages. MIT, Cambridge
2. Hockett CF (1960) The origin of speech. Sci Am 203:89–97
3. Skinner BF (1953) Science and human behavior. Macmillan, New York
4. Chomsky AN (1959) On certain formal properties of grammars. Inf Control 2:137–167
5. Pinker S, Bloom P (1990) Natural language and natural selection. Behav Brain Sci 13(4): 707–784
6. Pinker S (1995) The language instinct: the new science of language and mind. Penguin Books, Middlesex
7. Busch A, Boles WW, Sridharan S (2005) Texture for script identification. IEEE Trans Pattern Anal Mach Intell 27:1720–1732
8. Nakayama T, Spitz AL (1993) European language determination from image. In: Proceedings of the 2nd international conference on document analysis and recognition, Tsukuba, pp 159–162
9. Hochberg J, Bowers K, Cannon M, Kelly P (1999) Script and language identification for handwritten document images. Int J Doc Anal Recognit 2:45–52
10. Sibun P, Spitz AL (1994) Language determination: natural language processing from scanned document images. In: Proceedings of the applied natural language processing, Stuttgart, pp 15–21
11. Beesley KR (1988) Language identifier: a computer program for automatic natural-language identification of on-line text. In: Language at crossroads: proceedings of the 29th annual conference of the American Translators Association, Seattle, pp 47–54
12. Cavner WB, Trenkle JM (1994) N-gram based text categorization. In: Proceedings of the third annual symposium of document analysis and information retrieval, Las Vegas, pp 161–169
13. Cole RA, Mariani J, Uszkoreit H, Zaenen A, Zue V (eds) (1997) Survey of the state of the art in human language technology. Cambridge University Press, Cambridge
14. Hays J (1993) Language identification using two and three-letter cluster. Technical report, School of Computer Studies, University of Leeds
15. Ingle NC (1976) A language identification table. Inc Linguist 15:98–101
16. Mathusamy YK, Barnard E (1994) Reviewing automatic language identification. IEEE Signal Process Mag 11:33–41
17. Souter C, Churcher G, Hayes J, Hughes J, Johnson S (1994) Natural language identification using corpus-based models. In: Lauridsen K, Lauridsen O (guest eds) Hermes J Linguist 13:183–203
18. Majumder P, Mitra M, Chaudhuri BB (2002) N-gram: a language independent approach to IR and NLP. In: Proceedings of the international conference on universal knowledge and language, Goa, 25–29 Nov 2002
19. Dash NS (2011) A descriptive study of the modern Bengali script. Lambert Academic: Saarbrucken

20. Ghosh D, Dube T, Shivaprasad AP (2010) Script recognition-a review. IEEE Trans PAMI 32(12):2142–2161
21. Pal U, Roy PP, Tripathy N, Llados J (2010) Multi-oriented Bangla and Devnagari text recognition. Pattern Recognit 43:4124–4136
22. Spitz L (1990) Multilingual document recognition. In: Furuta R (ed) Electronic publishing, document manipulation, and typography. Cambridge University Press, Cambridge/ New York/Melbourne, pp 193–206
23. Spitz AL (1994) Text characterization by connected component transformation. In: Proceedings of SPIE, document recognition, San Jose, vol 2181, pp 97–105
24. Spitz L (1997) Determination of the script and language content of document images. IEEE Trans Pattern Anal Mach Intell 19:235–245
25. Hochberg J, Kelly P, Thomas T, Kerns L (1997) Automatic script identification from document images using cluster-based templates. IEEE Trans Pattern Anal Mach Intell 19: 176–181
26. Hochberg J, Kerns L, Kelly P, Thomas T (1995) Automatic script identification from images using cluster-based templates. In: Proceedings of the 3rd international conference on document analysis and recognition, Montreal, pp 378–381
27. Tan TN (1998) Rotation invariant texture features and their use in automatic script identification. IEEE Trans Pattern Anal Mach Intell 20:751–756
28. Peake GS, Tan TN (1997) Script and language identification from document images. In: Proc. Eighth British Mach. Vision Conf., Essex, UK, vol 2, pp 230–233
29. Pam WM, Suen CY, Bui T (2005) Script identification using steerable Gabor filters. In: Proceedings of the 8th international conference on document analysis and recognition, Seoul, pp 883–887
30. Ding J, Lam L, Suen CY (1997) Classification of oriental and European scripts by using characteristic features. In: Proceedings of the 4th international conference on document analysis and recognition, Ulm, pp 1023–1027
31. Zhang T, Ding X (1999) Cluster-based bilingual script-segmentation and language identification. In: Character recognition and intelligent information processing, Tsinghua University, China, vol 6, pp 137–148
32. Lee DS, Nohl CR, Baird HS (1996) Language identification in complex, un-oriented and degraded document images. In: Proceedings of the IAPR workshop on document analysis and systems, Malvern, pp 76–98
33. Ablavsky V, Stevens M (2003) Automatic feature selection with applications to script identification of degraded documents. In: Proceedings of the 7th international conference on document analysis and recognition, Edinburgh, pp 750–754
34. Elgammal M, Ismail MA (2001) Techniques for language identification for hybrid Arabic–English document images. In: Proceedings of the 6th international conference on document analysis and recognition, Seattle, pp. 1100–1104
35. Wood S, Yao X, Krishnamurthi K, Dang L (1995) Language identification for printed text independent of segmentation. In: Proceedings of the international conference on image processing, Washington, DC, pp 428–431
36. Tao Y, Tang YY (2001) Discrimination of oriental and Euramerican scripts using fractal feature. In: Proceedings of the 6th international conference on document analysis and recognition, Seattle, pp 1115–1119
37. Pal U, Chaudhuri BB (1999) Script line separation from Indian multi-script documents. In: Proceedings of the 5th international conference on document analysis and recognition, Bangalore, pp 406–409
38. Pal U, Sinha S, Chaudhuri BB (2003) Multi-script line identification from Indian documents. In: Proceedings of the 7th international conference on document analysis and recognition, Edinburgh, pp 880–884
39. Chaudhuri S, Sheth R (1999) Trainable script identification strategies for Indian languages. In: Proceedings of the fifth international conference on document analysis and recognition, Bangalore, pp 657–660

40. Pal U, Chaudhuri BB (2001) Automatic identification Of English, Chinese, Arabic, Devnagari and Bangla script line. In: Proceedings of the sixth international conference on document analysis and recognition, Seattle, pp 790–794
41. Wang KY, Casey RG, Wahl FM (1982) Document analysis system. IBM J Res Dev 26:647–656
42. Pal U, Roy PP (2004) Multi-oriented and curved text lines extraction from Indian documents. IEEE Trans Syst Man Cybern B 34:1676–1684
43. Pal U, Chaudhuri BB (1997) Automatic separation of words in Indian multi-lingual multi-script documents. In: Proceedings of the fourth international conference on document analysis and recognition, Ulm, pp 576–579
44. Patil SB, Subbareddy NV (2002) Neural network based system for script identification in Indian documents. Sadhana 27:83–97
45. Dhanya D, Ramakrishna AG, Pati PB (2002) Script identification in printed bilingual documents. Sadhana 27:73–82
46. Dhanya D, Ramakrishna AG (2002) Script identification in printed bilingual documents. In: Proceedings of the document analysis and systems, Princeton, pp 13–24
47. Manthalkar R, Biswas PK, An automatic script identification scheme for Indian languages. www.ee.iitb.ac.in/uma/~ncc2002/proc/NCC-2002/pdf/n028.pdf.
48. Mantas J (1986) An overview of character recognition methodologies. Pattern Recognit 19:425–430
49. Roy K, Pal U, Chaudhuri BB (2005) A system for neural network based word-wise handwritten script identification for Indian postal automation. In: Second international conference on intelligent sensing and information processing and control, Mysore, pp 581–586
50. Lam L, Ding J, Suen CY (1998) Differentiating between oriental and European scripts by statistical features. Int J Pattern Recognit Artif Intell 12(1):63–79
51. Hochberg J, Cannon M, Kelly P, White J (1997) Page segmentation using script identification vectors: a first look. In: Proceedings of the 1997 symposium on document image understanding technology, Annapolis, pp 258–264
52. Pal U, Chaudhuri BB (2002) Identification of different script lines from multi-script documents. Image Vis Comput 20(13–14):945–954
53. Padma MC, Nagabhushan P (2003) Identification and separation of text words of Kannada, Hindi and English languages through discriminating features. In: Proceedings of the 2nd Indian conference on document analysis and recognition, Mysore, India, pp 252–260
54. Jawahar CV, Pavan Kumar MNSSK, Ravi Kiran SS (2003) A bilingual OCR for Hindi–Telugu documents and its applications. In: Proceedings of the international conference on document analysis and recognition, Edinburgh, Aug 2003 pp 408–412
55. Chanda S, Sinha S, Pal U (2004) Word-wise English Devnagari and Oriya script identification. In: Sinha RMK, Shukla VN (eds) Speech and language systems for human communication. Tata McGraw-Hill, New Delhi, pp. 244–248
56. Jain AK, Zhong Y (1996) Page segmentation using texture analysis. Pattern Recognit 29(5):743–770
57. Singhal V, Navin N, Ghosh D (2003) Script-based classification of handwritten text documents in a multilingual environment. In: Proceedings of the 13th international workshop on research issues in data engineering–multilingual information management, Hyderabad, pp 47–53
58. Joshi GD, Garg S, Sivaswamy J (2006) Script identification from Indian documents. In: Proceedings of the IAPR international workshop document analysis systems, Feb 2006, Nelson, New Zealand, pp 255–267
59. Ma H, Doermann D (2003) Gabor filter based multi-class classifier for scanned document images. In: Proceedings of the international conference on document analysis and recognition, Aug 2003, Edinburgh, Scotland, pp 968–972
60. Jaeger S, Ma H, Doermann D (2005) Identifying script on word-level with informational confidence. In: Proceedings of the international conference on document analysis and recognition, Aug/Sept 2005, vol 1, pp 416–420
61. Chanda S, Pal U, Terrades OR (2009) Word-wise Thai and Roman script identification. ACM Trans Asian Lang Inf Process 8(11):1–21

62. Zhou L, Lu Y, Tan CL (2006) Bangla/English script identification based on analysis of connected component profiles. In: Proceedings of the 7th international workshop on document analysis and systems, Nelson, pp 243–254
63. Dhandra BV, Nagabhushan P, Hangarge M, Hegadi R, Malemath VS (2006) Script identification based on morphological reconstruction in document images. In: Proceedings of the international conference on pattern recognition (ICPR'06), Hong Kong, pp 950–953
64. Hou HS (1983) Digital document processing. Wiley, New York
65. Zramdini A, Ingold R (1993) Optical font recognition from projection profiles. Electron Publi 6(3):249–260
66. Zhu Y, Tan T, Wang Y (2001) Font recognition based on global texture analysis. IEEE Trans PAMI 23(10):1192–1200
67. Zramdini A, Ingold R (1998) Optical font recognition using typographical features. IEEE Trans PAMI 20(8):887–882
68. Khoubyari S, Hull JJ (1996) Font and function word identification in document recognition. Comput Vis Image Underst 63(1):66–74
69. Jeong CB, Kwag HK, Kim SH, Kim JS, Park SC (2003) Identification of font styles and typefaces in printed Korean documents. In: Sembok TMT et al (eds) ICADL 2003, Kuala Lumpur. LNCS 2911, pp 666–669
70. http://en.wikipedia.org/wiki/Languages_of_India
71. Sharma N., Chanda S, Pal U, Blumenstein U (2013) Word-wise script identification from video frames, 12th International conference on document analysis and recognition, Washington DC, USA pp 867–871
72. Morris RA (1988) Classification of digital typefaces using spectral signatures. Pattern Recognit 25(8):869–876
73. Anigbogu JCh (1992) Reconnaissance de Textes Imprimés Multifontes à l'Aide de Modèles Stochastiques et Métriques. Ph.D. dissertation, Université de Nancy I
74. Abuhaiba ISI (2003) Arabic font recognition based on templates. Int Arab J Inf Technol 1:33–39
75. http://www.ntchosting.com/multimedia/font.html
76. Chanda S, Pal U, Franke K (2012) Font identification: in context of an Indic script. In: Proceedings of the 21st international conference on pattern recognition (ICPR), Tsukuba, pp 1655–1658

Further Reading

Greenberg [1] presents a classification of world languages based on their typological features. Skinner [3] provides ideas about how human behaviors may be modeled after animal behavior controlled by some operant conditioning (this has been criticized by Chomsky in his early works). Reading Chomsky is mandatory for understanding the generative aspects of language. Pinker [6] presents an excellent narration about the growth and development of natural language as a part of natural selection – a theory based on the Darwinian model of natural selection. Cole, Mariani, Uszkoreit, Zaenen, and Zue [13] present an informative sketch on the state of the art in human language technology from the perspective of machine learning and knowledge representation. Pal and Chaudhuri [52] propose an effective strategy for identification of different script lines from multiscript documents. For more details about script identification, we refer to the paper [20]. The book by Dash [19] presents a complete picture about the form and function of characters and other orthographic symbols used in the modern Bengali script. Jain and Zhong [56] present a good model for page segmentation using a texture analysis system. The book by Hou [64] is a basic text on digital document processing and may be used for initial reading. Zhu, Tan, and Wang [66] propose a good method for font recognition based on global texture analysis.