REVIEW ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Construction of Alphabetic Character Recognition Systems: A Review

Hamsa D. Majeed*, Goran Saman Nariman

Department of Information Technology, College of Science and Technology, University of Human Development, Kurdistan Region, Iraq

ABSTRACT

Character recognition (CR) systems were attracted by a massive number of authors' interest in this field, and lot of research has been proposed, developed, and published in this regard with different algorithms and techniques due to the great interest and demand of raising the accuracy of the recognition rate and the reliability of the presented system. This work is proposed to provide a guideline for CR system construction to afford a clear view to the authors on building their systems. All the required phases and steps have been listed and clarified within sections and subsections along with detailed graphs and tables beside the possibilities of techniques and algorithms that might be used, developed, or merged to create a high-performance recognition system. This guideline also could be useful for readers interested in this field by helping them extract the information from such papers easily and efficiently to reach the main structure along with the differences between the systems. In addition, this work recommends to researchers in this field to comprehend a specified categorical table in their work to provide readers with the main structure of their work that shows the proposed system's structural layout and enables them to easily find the information and interests.

Index Terms: Optical Character Recognition, Script Identification, Document Analysis, Character Recognition, Multi-Script Documents

1. INTRODUCTION

In recent decades, many studies have demonstrated the ability of the machine to examine the environment and learn to distinguish patterns of interest from their background and make reliable and feasible decisions regarding the categories of the patterns. With huge volumes of data to be dealt with and through years of research, the design of approaches based on character recognition (CR) remains an ambiguous goal. Various frameworks employed machine learning approaches which have been most comprehensively studied and applied to a large number of systems that are essential in building a high-accuracy recognition system, CR is among the most well-known techniques and methods that make use of such artificial intelligence which have received attention increasingly. Moreover, in various application domains, ranging from computer vision to cybersecurity, character classifiers have shown splendid performance [1]-[3].

The application of CR is concerned with several fields of research. Through those numerous applications, there is no single approach for recognition or classification that is optimal and that motivates the researchers to explore multiple methods and approaches to employ. In addition, a combination of several techniques and classifiers is popped to the surface to serve the same purpose. Due to the increased attention paid to CR-based applications, noticeably there are few comprehensive overviews and systematic mappings of

Access this article online

DOI: 10.21928/uhdjst.v7n1y2023.pp32-42

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Hamsa D. Majeed, Department of Information Technology, College of Science and Technology, University of Human Development, Kurdistan Region, Iraq. E-mail: hamsa.al-rubaie@uhd.edu.iq

Received: 09-11-2022 Accepted: 07-02-2023 Published: 18-02-2023

CR applications design. Instead, the existing reviews explore in detail a specific domain, technique, or system focusing on the algorithms and methodology details [4], [5].

While starting investigations in this field, a big space of confusion appeared while diving into the details of each step in the recognition process due to the variety of paths that could be taken to reach the final goal and the pool of factors to be phished for that matter. That leads to the fact of considering an in-depth literature review as a requirement for surveying the possibility of using the techniques, approaches, or methodologies that are required for that phase of the recognition process among the others and deciding if they are suitable or not for that CR-based application.

The major aim of this study is to present the main path for the various kinds of approaches to be followed before diving into the details of the framework to be proposed by the meant research, Moreover, depending on each research field, there are options offered and categorized, techniques, and methods are presented and summarized from multiple perspectives all of which are investigated to answer the following queries:

1. Which language will be taken to recognize as input and what is a specified script writing style?
2. How can the data be acquired? Is it taken digitally (touch-screen, scanner, or another digital device) or uploaded from a non-digital source? In printed form by a keyboard or in handwritten form?
3. Which scale or level of detail is present in that set of data? Does the script have to be taken wholly or by a single character each time?
4. From which source could those data be collected? Is the preprocessing phase needed or not?
5. Generally, through which recognition process should invade for the optimal outcomes considering the previously chosen phases?

This work is structured to give the most suitable roadmap to the author of interest by presenting a systematic guideline to explore the multidisciplinary path starting from the script writing style the passing by the most suitable guide throughout the desired dataset characteristics (acquisition, granularity level, and the source of collected data), reaching to the script recognition process for the CR-based applications. Furthermore, this study uncovers the potential of CR applications among different domains and specifications by summarizing the purpose, methodologies, and application.

Thorough proofreading of several types of research including survey articles, the CR process has the same stations to stop by which could be sorted under some separated categories on specific factors and all those categories of any proposed system may have a stop in those main stations, that was an encouragement to make this study to highlight those main stations and present a guideline the researchers of interest by examining the detailed of sub-stations due to building CR system efficient to the author and understandable by the reader.

2. PROPOSED WALKTHROUGH GUIDELINE

The main goal of this study is to construct and design criteria for researchers working in the field of CR systems to observe when initiating research in both the practical and written parts. The following classifications and assortments are proposed, as shown in Fig. 1.

2.1. Script Writing System

From the linguistic point of view, nowadays, scripts used throughout the work have been broken down into six script classes, each of which can be used in one or more languages [6], [7]. Furthermore, in the context of CR, the investigations of the script character characteristics and structural properties, the script-written system has been classified under six classes. Different classes may contain the same language scripts [8], [9], [10]. Fig. 2 illustrates the classification of the script writing system.

2.1.1. Logographic system

The oldest kind of writing system is a logographic writing system; it is also called an ideogram as well, which employs symbols to depict a whole word or morpheme. The most well-known logographic script is Chinese, but logograms

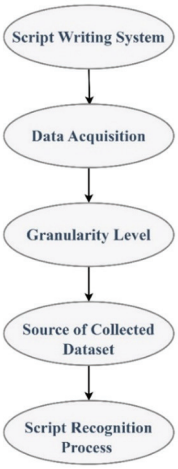

Fig. 1. General assortments of the CR system.

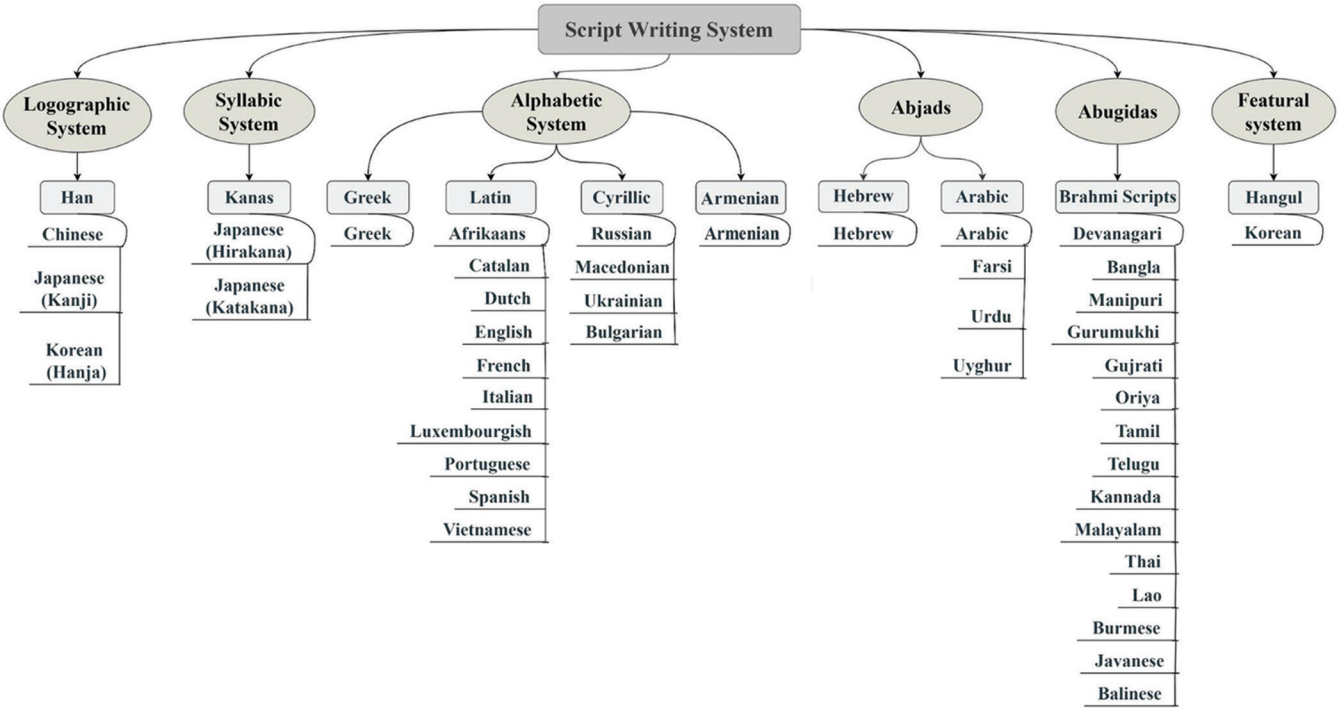

Fig. 2. Script Written System classifications.

such as numbers and the ampersand are found in almost all languages. An ideographic writing system typically has thousands of characters. Thus, the recognition process of this kind of script is still a challenging and fascinating topic for researchers. Han is the only script family in this class that includes two more languages, namely, Japanese (Kanji) and Korean (Hanja). The interesting distinguishing point between Han and other languages is the text line written direction, which is either from top to bottom or left to right.

In literature, lots of research can be found on handwritten CR in these scripts, for instance [11]-[13] work on Chinese, Japanese (Kanji), and Korean (Hanja), respectively. The accuracy rates for the scripts based on the aforementioned references are 99.39%, 99.64%, and 86.9%, respectively.

2.1.2. Syllabic system

Every written sign in a syllabic system, such as the one used in Japanese, corresponds to a phonetic sound or syllable. Kanas, which are divided into two types - Hirakana and Katakana - represent Japanese syllables. The Japanese script combines logographic Kanji and syllabic Kanas, as mentioned in the previous subsection. The Kanas has a similar visual appearance to the Chinese, with the exception that the Kanas has a lower density than the Chinese.

A lot of recognition progress can be found in the literature for both Hirakana and Katakana. Examples of excellent achievements in recognition accuracy rate are contributed in [11] for both Hirakana and Katakana, which are 98.83% and 98.19%, respectively.

2.1.3. Alphabetic system

Each consonant and vowel have a distinct symbol in the alphabetic writing system, which is used to write the languages classified under this written system. Segmental systems are another name for alphabets. To represent spoken language, these systems mix a small number of characters called letters. Letters are meant to represent certain phonemes. Greece is where the alphabet was first used, and it later expanded around the world, especially in Europe and a part of Asia as well [14], proposed a system for Ancient Greek CR that achieved an accuracy rate of 96%. Latin, Cyrillic, and Armenian also belong to this system.

There are numerous languages that use the Latin alphabet, commonly known as the Roman script, with differing degrees of alteration. It is utilized to write in a wide range of European languages, including English, French, Italian, Portuguese, Spanish, German, and others. The interested authors of Latin languages presented their ideas in terms of the recognition system for the different Latin languages,

for instance, Afrikaans 98.53% [15], Catalan 91.97% [16], Dutch 95.5% [17], English 98.4% [18], French 93.6% [19], Italian 92.47% [20], Luxembourgish (87.55 ± 0.24)% [21], Portuguese 93% [22], Spanish 97.08% [23], Vietnamese 97% [24], and German 99.7% [25].

Cyrillic has a separate letter set but is still relatively comparable to Latin. The Cyrillic writing system has been adopted by certain Asian and Eastern European languages, including Russian, Bulgarian, Ukrainian, and Macedonian, where the recognition rate is recorded for them as follows: Russian 83.42% [26], Bulgarian 89.4% [27], Ukrainian 95% [28], and Macedonian 93% [29].

Finally, the Armenian written system, this language classified as an Indo-European language belonging to an independent branch of which it is the only member recent CR system for this language scored 89.95% [30].

2.1.4. Abjads

When the words have a writing pattern from right to left along with text line, written in a repetition of consonants that are close together leaving the vowel sounds to be inferred by the reader, and have cursive long strokes consisting of few dots, then you are looking at Abjads writing system. It is unlike most other scripts in the world but it is similar to the alphabetic system unless it has symbols for consonantal sounds only. These unique features make the process of script identification for Abjads relatively simpler compared to other scripts, particularly because of the long cursive strokes with dots and the right-to-left writing direction, making it easier for recognition systems in pen computing.

Arabic and Hebrew are considered the major categories of the Abjads writing system. There are some other scripts of Arabic origin, such as Farsi (Persian), Urdu, and Uyghur. A lot of approaches had been proposed for identifying Abjad-based scripts, they used the long main stroke along with the cursive appearance yielding from conjoined words for Arabic. Meanwhile, the more uniform strokes in length and discrete letters were the main dependent features of Hebrew script recognition. According to the latest survey for Arabic recognition systems [31], the highest accuracy score is 99.98%, while recorded 97.15% for Hebrew [32]. In Farsi, Urdu, and Uyghur, the highest accuracies achieved are 99.45%, 98.82%, and 93.94%, respectively [33]-[35].

2.1.5. Abugidas

It is a writing script primarily based on a consonant letter and secondary vowel notation. They are sharing with alphabetic systems the property of combining characters writing styles within the text line. It belongs to the Brahmic family of scripts which is can be expressed in two groups:

1. Original Brahmi script: This northern group deployed in Devnagari, Bangla (Bengali), Manipuri, Gurumukhi, Gujrati, and Oriya languages. The most recent survey papers for the CR systems of this group come up with the highest recognition rate of 99% for Devnagari, 99.32% for Bangla (Bengali), 98.70% for Manipuri, 99.3% for Gurumukhi, 98.78% for Gujrati, and 96.7% for Oriya [36]-[38].
2. Derived from Brahmi: Look quite different from the northern group and used in:
    a. South India: Tamil, Telugu, Kannada, and Malayalam, where the highest accuracy of the mentioned language for recognition matter was for Tamil 98.5%, Telugu 98.6%, Kannada 92.6%, and Malayalam 98.1% [39].
    b. Southeast Asia: Thai, Lao, Burmese, Javanese, and Balinese, the languages of this group have achieved the highest validation rate where Thai, Lao, and Burmese attained 92.1%, 92.41%, and 96.4% while Javanese and Balinese gained 97.7% and 97.53%, respectively [40]-[43].

2.1.6. Featural system

This form of writing system is significantly represented by symbols or characters, the main language is Korean which is described as less complex and less dense compared to Chinese and Japanese, it is represented by mixing logographic Hanja and featural Hangul, the highest scored accuracy rate for Korean was 97.07% [44].

As a summarization of all the findings in this section, Table 1 illustrates the classifications of the languages with the highest accuracy recorded so far.

2.2. Data Acquisition

The next step for the author after selecting which language to work on is to decide which writing style will be chosen for recognition, this step is considered one of the fixed and essential phases in all the recognition studies and research, reaching this phase requires the knowledge of how to start acquiring data to be fed into the recognition system, the answer simply starts with defining the writing style, here the author has two options either printed script or handwritten script.

After making the decision, the acquisition tools are required either offline tools or online. In this section, a guideline is

TABLE 1: Summarization of languages with their recent highest accuracy rate

Script writing system	Main language	Sub-language	Accuracy rate (%)
Logographic system	Han	Chinese	99.39
		Japanese (Kanji)	99.64
		Korean (Hanja)	86.9
Syllabic system	Kanas	Japanese (Hirakana) a	98.83
		Japanese (Katakana)	98.19
Alphabetic system	Greek	Greek	96
	Latin	Afrikaans	98.53
		Catalan	91.97
		Dutch	95.5
		English	98.4
		French	93.6
		Italian	92.47
		Luxembourgish	(87.55±0.24)
		Portuguese	83
		Spanish	97.08
		Vietnamese	97
		German	99.7
	Cyrillic	Russian	83.42
		Bulgarian	89.4
		Ukrainian	95
		Macedonian	93
	Armenian	Armenian	89.95
Abjads	Hebrew	Hebrew	97.15
	Arabic	Arabic	99.98
		Farsi	99.45
		Urdu	98.82
		Uighur	93.94
Abugidas	Brahmi	Devnagari	99
		Bangla (Bengali)	99.32
		Manipuri	98.70
		Gurumukhi	99.3
		Gujrati	98.78
		Oriya	96.7
		Tamil	98.5
		Telugu	98.6
		Kannada	92.6
		Malayalam	98.1
		Thai	92.1
		Lao	92.41
		Burmese	96.4
		Javanese	97.7
		Balinese	97.53
Featural system	Korean	Korean	97.07

proposed and could be followed to help make those decisions as Fig. 3 shows.

2.2.1. Printed character

Those characters are produced as a result of the process of producing using inked-type tools. In recognition systems of any language, the printed characters usually achieve a high recognition rate because it is considered in regular form, clean, have the same style, and have similar shapes and lines,

Fig. 3. Overview of data acquisition.

and that facilitates the learning operation and therefore raises the accuracy of recognition in the testing phase.

2.2.2. Handwriting character

When the process of forming letters of any language is done with the hand, rather than any typing device then the result is handwriting characters. Most of the authors that are interested in CR are employing handwriting characters as input to their approaches to prove the effectiveness and efficiency of their systems or techniques due to the complexity and impenetrability that come with the variety of the handwriting style and the use of tools besides the differences in lines and colors not to mention the irregular shapes and positions.

2.2.3. Online character

These characters are obtained from digital devices with a touch screen with/without a keyboard involved like a personal digital assistant, or mobile. Where screen sensors receive the switching of pushing and releasing the pen on the screen in addition to the pen tip movements over the screen.

2.2.4. Offline character

This kind of character is attended when image processing is involved by converting an input image (from a scanner or a camera) of text to character code which is aimed to be utilized by a text-processing application.

It is essential for the author to choose the correct combination of the writing style and the writing tool, as Fig. 3 illustrates there are three combinations to decide among them: offline-printed where the input of the CR system decided to be in offline mode with characters taken from the printed device rather than the offline-handwritten which taken from a human-hand in offline-mode already written on paper in a previous time while the online-handwritten fed as input to CR system instantly by hand through a touchable input device without a keyboard.

Some recent recognition systems are illustrated in Table 2 for several languages to show some authors' choices for the

language, writing style, and writing tool, and how their choices affect the accuracy rate for each mechanism. Furthermore, a comprehensive survey for online and offline handwriting recognition can be found in Plamondon and Srihari [45].

The outcomes of Table 2 show that most existing studies have focused on handwritten text, with fewer works attempting to classify or identify printed text. This is because of the high variance in handwriting styles across people and the poor quality of the handwritten text compared to printed text yields the fact that handwritten CR is more challenging than the printed one.

On the other hand, it is noticeable using offline as writing tool more than online ones this is due to in the online case, features can be extracted from both the pen trajectory and the resulting image, whereas in the offline case, only the image is available, so the offline recognition is observed as harder than online recognition.

2.3. Granularity Level of Documents

The third type of classification of character handwriting recognition is "Granularity Level of Documents," which describes the level of detailed information taken as initial input to the defined and proposed framework. This class could be split into five granularity levels as shown in Fig. 4, from a script page full of text to a single letter or symbol.

TABLE 2: Examples of recognition systems with different data acquisition mechanisms

Reference	Language	Writing style	Writing tool	Accuracy rate (%)
[46]	Arabic	Handwritten	Offline	99.93
[18]	English	Handwritten	Offline	98.4
[47]	English	Printed	Offline	98
[48]	English	handwritten	Online	93.0
[49]	Chinese	Handwritten	Online	98
[50]	Chinese	Handwritten	Offline	94.9
[13]	Chinese	Printed	Offline	99.39
[51]	Arabic	Printed	Offline	97.51
[52]	Arabic	Handwritten	Online	96

In the domain of CR, if the initial input into the OCR framework is not at a character level, the process of script identification must proceed until it gets to a single character. This procedure, known as "Segmentation," will be covered in the following subsection (3.5).

2.3.1. Document/page level

Document-level script is the most detailed granularity level, where the entire document is exposed to the script identification procedure at once. Following processing, the document is further broken down into pages, pieces of paragraphs, text lines, words, and finally characters to enable the recognition of the precise letter. Although some researchers discriminate between the script recognition process at the document and page levels, in general, the technical methodologies are very similar. Because of this, some researchers alternately refer to document-level and page-level script recognition.

Finding the text region on a page is the initial step in page-level script identification. It is possible to carry out this operation by separating the pages into text and non-text pieces [53]. Several pieces of research can be found in the literature for both offline-handwritten [54] and offline-printed [55].

After the page of the script has been identified, the process of the next level starts, which is paragraph or text block identification. It operates by dividing the entire page into equal-sized text blocks with several lines of content. Text blocks can have different sizes, and padding may be necessary if characters are on the edge of a text block [56]. Is an example of segmenting pages into pieces of text blocks.

2.3.2. Paragraph level

The text block is separated into lines. The white space between lines is typically used for text line segmentation. Lines of scrip are detected and segmented to be prepared for further segmentation processing. Both offline-handwritten [57] and offline-printed [58] line detection has been the subject of numerous studies in the literature.

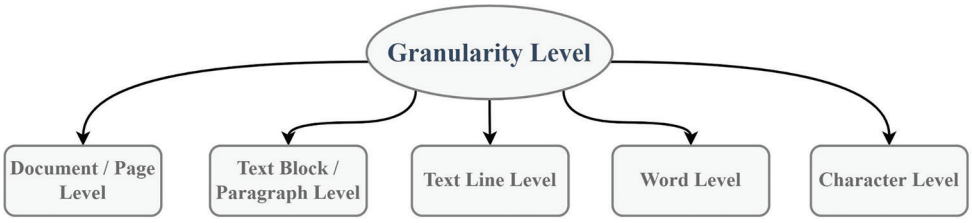

Fig. 4. Granularity level classification.

2.3.3. Text line level

A framework that gets a line of script as initial input needs segmentation processes to identify each word in the text. Therefore, word identification is needed. Text lines are divided into words; usually, the white space between text lines is used for this purpose. Numerous literary attempts have been made to address the difficulties encountered in this process. For instance, there might be noise, twisted words, missing or partial letters in words, words that are not available as straight text lines, etc. Some examples given in this topic of identifying words in a text line are [59] for offline-handwritten and [60] for offline-printed.

2.3.4. Word level

Character detection and segmentation are required since the initial input is a word. It usually works by combining properties from various characters to ensure the process. Several attempts have been made to improve accuracy and ensure that no character inside a word is missed. For instance, recently [61] used distinct strategies and achieved a satisfactory outcome.

2.3.5. Character level

Finally, there is no requirement for segmentation at the character level because the initial input into the proposed framework is already character. The character goes through preprocessing, which is followed by recognition procedures. In some circumstances, no preprocessing is required, as is the case when using a character public dataset. For instance [62], is an example of working at the character level with and without preprocessing, respectively.

In addition, to avoid confusion between granularity levels for identification/detection and recognition processes, it is worth mentioning that from the recognition standpoint, when the granularity level is text line level, it means that the text line is already known and the detection and segmentation into words and characters are needed. However, from the identification/detection point of view, it means that the identification and detection of text lines are working. Further details about these processes can be found in [10], [63].

2.4. Source of Collected Dataset

The essential component of any machine learning application is the dataset. That leads us to discuss this important phase of CR as the fourth classification named Source of Collected Dataset which is broken down into two categories as Fig. 5 illustrates:

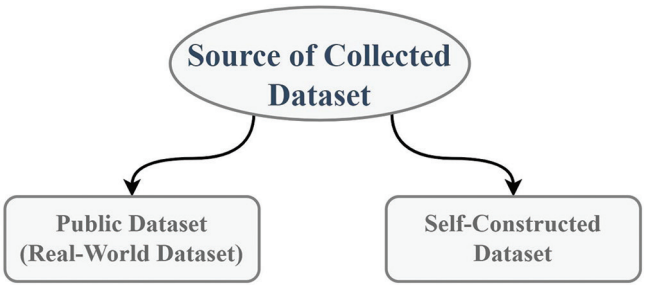

Fig. 5. Categories of collected dataset sources.

2.4.1. Public dataset (real-world dataset)

The term "public dataset" refers to a dataset saved in the cloud and made open to the public. MNIST, Keras, Kaggle, and others are examples. Almost all of the public datasets have been preprocessed, cleaned, and usually, in the case of character level, reshaped to 28 × 28 pixels and saved as CSV files. Many authors attain to use this source to skip the preprocessing step and focus more on the other steps and easily find opponents for the comparison issue of those who used the same data source with different techniques.

2.4.2. Self-constructed dataset

Is the dataset that the researchers create and prepare on their own depending on their techniques, it is an online or offline way of collection, this source of dataset is considered more challenging because the collected images are not processed at all in terms of resizing, denoising, colored, etc.

For a fair comparison, this kind of work better to be compared with studies that have done with a self-collected source of data, not with a public one that comes clean and processed. Researchers should be aware of the data to be collected and use the proper tools required to preprocess in a way that suits the technique used for recognition.

2.5. Script Recognition Process

The script recognition process (the implementable phase) is the fifth classification type of alphabet handwritten recognition framework. In an in-depth study of several research articles, including survey articles, we mainly focused on the phases that an OCR system needs to accomplish its recognition goal. Thus, we could conclude that four categories can be defined based on the number of phases in which the whole procedure of recognition comprises, as presented in Fig. 6.

In addition, commonly, script recognition is achieved by blending traditional image processing techniques with

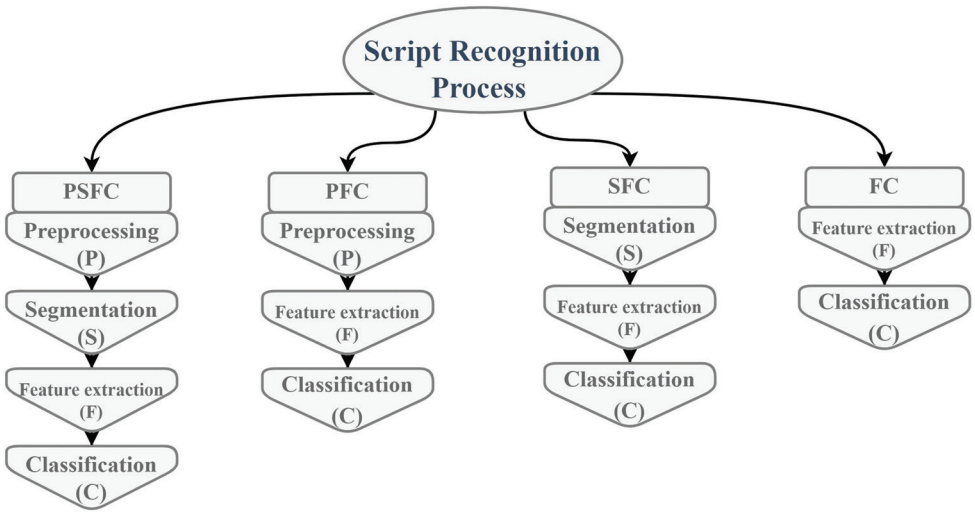

Fig. 6. Basic components of the script recognition processes.

image identification and recognition techniques. The recognition composition is formed from four primary phases, namely, Preprocessing (P), Segmentation (S), Feature Extraction (F), and Classification (C). The last two phases, Feature-Extraction and Classification, are the most common in the research. There is not any work without any of these two phases. The next few paragraphs will briefly outline them.

- Preprocessing (P) is a sequence of operations performed to intensify the input image. It is responsible for removing noise, resizing, thinning, contouring, transforming the dataset into a black-white format, edge detection, etc. Every single one of them can be performed with an appropriate technique
- Segmentation (S) performs the duty of obtaining a single character. The document processing follows a hierarchy; it starts from the whole page and ends with a single character. The required level of the hierarchy is a single character
- Feature extraction (F) is a mechanism in which each character is turned into a feature vector using specific algorithms for the extraction of the features, which is then fed into a classifier to determine which class it belongs to.
- The classification (C) phase is a decision-making process that uses the features extracted from the preceding step as input. And it decides what the final output is.

It is worth noticing that handwritten mathematical symbols and expressions recognition is out of our research scope. Therefore, we do not consider the two additional phases (Structural Analysis and Symbol Recognition) which are included in such works. More details can be found in Sakshi and Kukreja [64].

2.5.1. PSFC

The first category of the Script Recognition Process class can be called PSFC, which means all four phases have been utilized to achieve the goal as [65] describe.

2.5.2. PFC

The segmentation process is skipped in the second category, in most cases due to working on character level as initial input therefore no need for segmentation as presented in Parthiban *et al.* [66].

2.5.3. SFC

The third one is SFC as [67] proposal, where the preprocessing is missed because the entered data originally is clean and there is no preprocessing required.

2.5.4. FC

In the fourth and last category as illustrated in Gautam and Chai [68], the first two phases P and F are dismissed because the granularity level is letters, and the initial input data is originally clean. For instance, works utilizing public datasets such as MNIST [69] could be classified under this category.

3. EXAMPLES

This section is to illustrate some of the CR systems and gives a description of how to read their roadmap regarding their systems, by applying the proposed guideline, any paper in this field can be summarized in stages according to the

TABLE 3: Examples of the proposed framework of character recognition

Example 4 [48]		Example 5 [13]	
Classifications	**Nominated category**	**Classifications**	**Nominated category**
Script writing system	English	Script writing system	Chinese
Data acquisition	Online-handwritten	Data acquisition	Offline-printed
Granularity level of documents	Line level	Granularity level of documents	Character level
Source of the collected dataset	Public dataset	Source of the collected dataset	Self-constructed dataset
Script recognition process	FC	Script recognition process	PFC

author's choices and be easier to the reader to figure out the main stages and for the other authors to develop any desired CR system.

Some examples are presented here to show how the CR system can be summarized according to the proposed guideline, and resembling a table is suggested to be created in such a work to provide a comprehensive view of the proposed framework as a whole. It makes it easier for the reader to find the information they are searching for before going into depth. Table 3 provides two examples of how to present the suggested table. In addition, the following examples demonstrate how the systems may be constructed using the component chain:

1. [18] English → offline-handwritten → character level → self-constructed dataset → PFC
2. [46] Arabic → offline-handwritten → line level → public dataset → PSFC
3. [49] Chinese → online-handwritten → page level → public dataset → SFC
4. [48] English → online-handwritten → line level → public dataset → FC
5. [13] Chinese → offline-printed → character level → self-constructed dataset → PFC

4. CONCLUSION

CR stepped ahead as an eminent topic of research. Exhaustive studies continuously presented CR of different languages with various algorithms that were developed to increase the reliability of these characters for accurate recognition. A guideline for the construction CR system has been proposed for the authors in this field to overcome the unclear presentation and expressing ideas in such a domain of science. Almost all the required steps have been shown and demonstrated by graph and table to be used in such works in CR Domaine for more clarity for the authors to margin their scope. It is also for the readers, as well, to directly recognize the used technique through in-text reading and then move forward to the details afterward. Through reading this guideline, the authors will be able to order their thoughts and build their recognition system smoothly and effectively especially for the new authors in this field, as for readers after reading this work they will have the ability to analyze other research in the relative fields and extract information easily from other works of interest, for the seekers of new ideas or merging techniques, this guideline is suitable to help to determine the exact part of recognition system to be studied or compared with. Saving time, effort, and thoughts orienting for other authors or readers was one of the essential aims of this work.

REFERENCES

[1] M. Paolanti and E. Frontoni. "Multidisciplinary pattern recognition applications: A review". *Computer Science Review*, vol. 37, pp. 100276, 2020.

[2] M. Kawaguchi, K. Tanabe, K. Yamada, T. Sawa, S. Hasegawa, M. Hayashi and Y. Nakatani. "Determination of the Dzyaloshinskii-Moriya interaction using pattern recognition and machine learning". *npj Computational Materials*, vol. 7, no. 1, 2021.

[3] B. Biggio and F. Roli. "Wild patterns: Ten years after the rise of adversarial machine learning". *Pattern Recognition*, vol. 84, pp. 317-331, 2018.

[4] T. S. Gorripotu, S. Gopi, H. Samalla, A. V. Prasanna and B. Samira. "Applications of Computational Intelligence Techniques for Automatic Generation Control Problem-a Short Review from 2010 to 2018." In: Computational Intelligence in Pattern Recognition. Springer Singapore, Singapore, 2020, pp. 563-578.

[5] M. I. Sharif, J. P. Li, J. Naz and I. Rashid. "A comprehensive review on multi-organs tumor detection based on machine learning". *Pattern Recognition Letters*, vol. 131, pp. 30-37, 2020.

[6] A. Nakanishi. "Writing Systems of the World: Alphabets, Syllabaries, Pictograms". Charles E. Tuttle Co., United States, 1980.

[7] F. Coulmas. "The Blackwell Encyclopedia of Writing Systems". Blackwell, London, England, 1999.

[8] D. Sinwar, V. S. Dhaka, N. Pradhan and S. Pandey. "Offline script recognition from handwritten and printed multilingual documents: A survey". *International Journal on Document Analysis and Recognition*, vol. 24, no. 1-2, pp. 97-121, 2021.

[9] D. Ghosh, T. Dube and A. P. Shivaprasad. "Script recognition--a review". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142-2161, 2010.

[10] K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo and I. Yibulayin. "Script identification of multi-script documents: A survey". *IEEE*

Access, vol. 5, pp. 6546-6559, 2017.

[11] C. Tsai. "Recognizing handwritten Japanese Characters using Deep Convolutional Neural Networks". University of Stanford in Stanford, California, pp. 405-410, 2016.

[12] S. Purnamawati, D. Rachmawati, G. Lumanauw, R. F. Rahmat and R. Taqyuddin. "Korean letter handwritten recognition using deep convolutional neural network on android platform". *Journal of Physics Conference Series*, vol. 978, no. 1, p. 012112, 2018.

[13] Y. Q. Li, H. S. Chang and D. T. Lin. "Large-scale printed Chinese character recognition for ID cards using deep learning and few samples transfer learning". *Applied Sciences*, vol. 12, no. 2, p. 907, 2022.

[14] B. Robertson and F. Boschetti. "Large-scale optical character recognition of ancient Greek". *Mouseion Journal of the Classical Association of Canada*, vol. 14, no. 3, pp. 341-359, 2017.

[15] J. Hocking and M. Puttkammer. "Optical Character Recognition for South African languages". In: 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016.

[16] A. Fornes, V. Romero, A. Baró, J. I. Toledo, J. A. Sánchez, E. Vidal, J. Lladós. "ICDAR2017 Competition on Information Extraction in Historical Handwritten Records". In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017.

[17] H. van Halteren and N. Speerstra. "Gender recognition on Dutch tweets". *Computational Linguistics in the Netherlands Journal*, vol. 4, pp. 171-190, 2019.

[18] H. D. Majeed and G. S. Nariman. "Offline handwritten English alphabet recognition (OHEAR)". *UHD Journal of Science and Technology*, vol. 6, no. 2, pp. 29-38, 2022.

[19] K. Todorov and G. Colavizza. "An Assessment of the Impact of OCR Noise on Language Models". In: Proceedings of the 14th International Conference on Agents and Artificial Intelligence, 2022.

[20] M. Del Buono, L. Boatto, V. Consorti, V. Eramo, A. Esposito, F. Melcarne and M. Tucci. "Recognition of Handprinted Characters in Italian Cadastral Maps". In: Character Recognition Technologies. SPIE Proceedings, 1993. vol. 1906, pp. 89-99.

[21] R. Barman, M. Ehrmann, S. Clematide, S. A. Oliveira and F. Kaplan, "Combining visual and textual features for semantic segmentation of historical newspapers. *Journal of Data Mining and Digital Humanities*, 2021.

[22] F. Lopes, C. Teixeira and H. G. Oliveira. "Comparing different methods for named entity recognition in Portuguese neurology text". *Journal of Medical Systems*, vol. 44, no. 4, p. 77, 2020.

[23] N. Alrasheed, P. Rao and V. Grieco. "Character Recognition of seventeenth-century Spanish American notary records using deep learning". *Digital Humanities Quarterly*, vol. 15, no. 4, 2021.

[24] T. Q. Vinh, L. H. Duy and N. T. Nhan. "Vietnamese handwritten character recognition using convolutional neural network". *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 276-283, 2020.

[25] A. Chaudhuri, K. Mandaviya, P. Badelia and S. K. Ghosh. "Optical character recognition systems for German language." In: Optical Character Recognition Systems for Different Languages with Soft Computing. Cham, Springer International Publishing, 2017, pp. 137-164.

[26] D. Gunawan, D. Arisandi, F. M. Ginting, R. F. Rahmat and A. Amalia. "Russian character recognition using self-organizing map". *Journal of Physics*: *Conference Series*, vol. 801, p. 012040, 2017.

[27] G. Georgiev, P. Nakov, K. Ganchev, P. Osenova and K. I. Simov. "Feature-rich Named Entity Recognition for Bulgarian using Conditional Random Fields". In: Proceedings of the International Conference RANLP-2009. arXiv [cs.CL], 2021.

[28] A. Radchenko, R. Zarovsky and V. Kazymyr, "Method of Segmentation and Recognition of Ukrainian License Plates". In: 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF), 2017.

[29] M. Gjoreski, G. Zajkovski, A. Bogatinov, G. Madjarov, D. Gjorgjevikj and H. Gjoreski. "Optical Character Recognition Applied on Receipts Printed in Macedonian Language". In: International Conference on Informatics and Information Technologies (CIIT), 2014.

[30] T. Ghukasyan, G. Davtyan, K. Avetisyan and I. Andrianov. "PioNER: Datasets and Baselines for Armenian Named Entity Recognition". In: 2018 Ivannikov Ispras Open Conference (ISPRAS), 2018.

[31] N. Alrobah and S. Albahli. "Arabic handwritten recognition using deep learning: A survey". *Arabian Journal for Science and Engineering*, 2022.

[32] O. Keren, T. Avinari, R. Tsarfaty and O. Levy, "Breaking Character: Are Subwords Good Enough for MRLs after all?" arXiv [cs.CL], 2022.

[33] Y. A. Nanehkaran, D. Zhang, S. Salimi, J. Chen, Y. Tian and N. Al-Nabhan. "Analysis and comparison of machine learning classifiers and deep neural networks techniques for recognition of Farsi handwritten digits". *Journal of Supercomputing*, vol. 77, no. 4, pp. 3193-3222, 2021.

[34] D. Rashid and N. Kumar Gondhi. "Scrutinization of Urdu handwritten text recognition with machine learning approach". In: Communications in Computer and Information Science. Cham, Springer International Publishing, 2022, pp. 383-394.

[35] Y. Wang, H. Mamat, X. Xu, A. Aysa and K. Ubul. Scene Uyghur text detection based on fine-grained feature representation". *Sensors* (*Basel*), vol. 22, no. 12, p. 4372, 2022.

[36] S. Sharma and S. Gupta. "Recognition of various scripts using machine learning and deep learning techniques-A review". In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021.

[37] P. D. Doshi and P. A. Vanjara. "A Comprehensive survey on Handwritten Gujarati Character and its Modifier Recognition Methods". In: Information and Communication Technology for Competitive Strategies (ICTCS 2020). Springer Singapore, Singapore, 2022, pp. 841-850.

[38] M. R. Haque, M. G. Azam, S. M. Milon, M. S. Hossain, M. A. A. Molla and M. S. Uddin. "Quantitative Analysis of deep CNNs for Multilingual Handwritten Digit Recognition". In: Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2021, pp. 15-25.

[39] H. Singh, R. K. Sharma and V. P. Singh. "Online handwriting recognition systems for Indic and non-Indic scripts: A review". *Artificial Intelligence Review*, vol. 54, no. 2, pp. 1525-1579, 2021.

[40] L. Saysourinhong, B. Zhu and M. Nakagawa. "Online handwritten Lao character recognition by MRF". *IEICE Transactions on Information and Systems*, vol. E95.D, no. 6, pp. 1603-1609, 2012.

[41] C. S. Lwin and W. Xiangqian. "Myanmar Handwritten Character Recognition from Similar Character Groups using K-means and Convolutional Neural Network". In: 2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE), 2020.

[42] M. A. Rasyidi, T. Bariyah, Y. I. Riskajaya and A. D. Septyani. "Classification of handwritten Javanese script using random forest

algorithm". *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1308-1315, 2021.

[43] I. W. A. Darma and N. K. Ariasih. "Handwritten Balinesse Character Recognition using K-Nearest Neighbor". INA-Rxiv, 2018.

[44] J. Park, E. Lee, Y. Kim, I. Kang, H. I. Koo and N. I. Cho. "Multi-lingual optical character recognition system using the reinforcement learning of character segmenter". *IEEE Access*, vol. 8, pp. 174437-174448, 2020.

[45] R. Plamondon and S. N. Srihari. "Online and off-line handwriting recognition: A comprehensive survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.

[46] N. S. Guptha, V. Balamurugan, G. Megharaj, K. N. A. Sattar and J. D. Rose, "Cross lingual handwritten character recognition using long short term memory network with aid of elephant herding optimization algorithm". *Pattern Recognition Letters*, vol. 159, pp. 16-22, 2022.

[47] G. S. Katkar and M. V Kapoor. "Performance analysis of structure similarity algorithm for the recognition of printed cursive English alphabets". *International Journal of Scientific Research in Science and Technology*, vol.8, no.5, pp. 555-559, 2021.

[48] S. Tabassum, N. Abedin, M. M. Rahman, M. M. Rahman, M. T. Ahmed, R. I. Maruf and A. Ahmed. "An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery". *Scientific Reports*, vol. 12, no. 1, p. 3601, 2022.

[49] D. H. Wang, C. L. Liu, J. L. Yu and X. D. Zhou. "CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters". In: 2009 10th International Conference on Document Analysis and Recognition, 2009.

[50] T. Q. Wang, X. Jiang and C. L. Liu. "Query pixel guided stroke extraction with model-based matching for offline handwritten Chinese characters". *Pattern Recognition*, vol. 123, p. 108416, 2022.

[51] A. Qaroush, B. Jaber, K. Mohammad, M. Washaha, E. Maali and N. Nayef. "An efficient, font independent word and character segmentation algorithm for printed Arabic text". *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 1, pp. 1330-1344, 2022.

[52] K. M. M. Yaagoup and M. E. M. Musa. "Online Arabic handwriting characters recognition using deep learning". *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 9, no. 10, pp. 83-92, 2020.

[53] P. B. Pati, S. Sabari Raju, N. Pati and A. G. Ramakrishnan. "Gabor Filters for Document Analysis in Indian Bilingual Documents." In: International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of, 2004, pp. 123-126

[54] S. M. Obaidullah, C. Halder, N. Das sand K. Roy. "Numeral script identification from handwritten document images". *Procedia Computer Science*, vol. 54, pp. 585-594, 2015.

[55] R. Bashir and S. Quadri. "Identification of Kashmiri Script in a Bilingual Document Image". In: 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), 2013.

[56] S. Manjula and R. S. Hegadi. "Identification and Classification of Multilingual Document using Maximized Mutual Information". In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[57] K. Roy, O. M. Sk, C. Halder, K. Santosh and N. Das. "Automatic line-level script identification from handwritten document images-a region-wise classification framework for Indian subcontinent". *Malaysian Journal of Computer Science*, vol. 31, no. 1, p. 10, 2016.

[58] G.S. Rao, M. Imanuddin and B. Harikumar. "Script Identification of Telugu, English and Hindi document image". *International Journal of Advanced Engineering and Global Technology*, vol. 2, no. 2, pp. 443-452, 2014.

[59] E. O. Omayio, I. Sreedevi and J. Panda. "Word Segmentation by Component Tracing and Association (CTA) Technique". *Journal of Engineering Research*, 2022.

[60] P. K. Singh, R. Sarkar and M. Nasipuri. "Offline script identification from multilingual Indic-script documents: A state-of-the-art". *Computer Science Review*, vol. 15-16, pp. 1-28, 2015.

[61] Y. Baek, D. Nam, S. Park, J. Lee, S. Shin, J. Baek, C. Y. Lee and H. Lee. "CLEval: Character-level Evaluation for Text Detection and Recognition Tasks". In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[62] K. J. Taher and H. D. Majeed. "Recognition of handwritten English numerals based on combining structural and statistical features". *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, vol. 21, no. 1, pp. 73-83, 2021.

[63] D. Sinwar, V. S. Dhaka, N. Pradhan and S. Pandey. "Offline script recognition from handwritten and printed multilingual documents: A survey". *International Journal on Document Analysis and Recognition*, vol. 24, no. 1-2, pp. 97-121, 2021.

[64] Sakshi and V. Kukreja. "A retrospective study on handwritten mathematical symbols and expressions: Classification and recognition". E*ngineering Applications of Artificial Intelligence*, vol. 103, p. 104292, 2021.

[65] N. Murugan, R. Sivakumar, G. Yukesh and J. Vishnupriyan. "Recognition of Character from Handwritten". In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1417-1419.

[66] R. Parthiban, R. Ezhilarasi and D. Saravanan. "Optical Character Recognition for English Handwritten Text using Recurrent Neural Network". In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020.

[67] H. Q. Ung, C. T. Nguyen, K. M. Phan, V. T. M. Khuong and M. Nakagawa. "Clustering online handwritten mathematical expressions". *Pattern Recognition Letters*, vol. 146, pp. 267-275, 2021.

[68] N. Gautam and S. S. Chai. "Zig-zag diagonal and ANN for English character recognition". *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.4, pp. 57-62, 2019.

[69] L. Deng. "The MNIST database of handwritten digit images for machine learning research [best of the web]". *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141-142, 2012.