A Detailed Analysis of Optical Character Recognition Technology

Article *in* International Journal of Applied Mathematics Electronics and Computers · December 2016

DOI: 10.18100/ijamec.270374

CITATIONS
23

READS
10,756

2 authors:

Karez Hamad
Soran University
1 PUBLICATION **23** CITATIONS
SEE PROFILE

Mehmet Kaya
Adiyaman University
10 PUBLICATIONS **53** CITATIONS
SEE PROFILE

International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org

Advanced Technology and Science

Original Research Paper

A Detailed Analysis of Optical Character Recognition Technology

Karez Abdulwahhab Hamad *[1], Mehmet Kaya [2]

Accepted 3rd September 2016

Abstract: In many different fields, there is a high demand for storing information to a computer storage disk from the data available in printed or handwritten documents or images to later re-utilize this information by means of computers. One simple way to store information to a computer system from these printed documents could be first to scan the documents and then store them as image files. But to re-utilize this information, it would very difficult to read or query text or other information from these image files. Therefore a technique to automatically retrieve and store information, in particular text, from image files is needed. Optical character recognition is an active research area that attempts to develop a computer system with the ability to extract and process text from images automatically. The objective of OCR is to achieve modification or conversion of any form of text or text-containing documents such as handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Therefore, OCR enables a machine to automatically recognize text in such documents. Some major challenges need to be recognized and handled in order to achieve a successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. Due to these challenges, characters sometimes may not be recognized correctly by computer system. In this paper we investigate OCR in four different ways. First we give a detailed overview of the challenges that might emerge in OCR stages. Second, we review the general phases of an OCR system such as pre-processing, segmentation, normalization, feature extraction, classification and post-processing. Then, we highlight developments and main applications and uses of OCR and finally, a brief OCR history are discussed. Therefore, this discussion provides a very comprehensive review of the state-of-the-art of the field.

Keywords: *OCR, OCR Challenges, OCR Phases, OCR Applications, OCR History.*

1. Introduction

It is natural and accustomed that we should demand to build and design machines that can recognize patterns. From automated optical character recognition to face recognition, fingerprint identification, speech recognition, DNA sequence identification and much more, it is clear that accurate and reliable pattern recognition by machine would be greatly useful.

Optical character recognition is an active research area that attempts to develop a computer system with the ability to extract and process text from images automatically. These days there is a huge demand for storing information to a computer storage disk from the data available in printed or handwritten documents to later re-utilize this information by means of computers. One simple way to store information to computer system from these paper documents could be to first scan the documents and then store them as image files. But to re-utilize this information, it would very difficult to read or query text or other information from these image files. Therefore a technique to automatically retrieve and store information, in particular text, from image files is needed. Of course, this is not a very trivial task. Some major challenges need to be laid out and handled in order to achieve a successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. Due to these challenges, characters sometimes may not be recognized correctly by computer system. Thus there is a need of mechanisms of character recognition to perform Document Image Analysis (DIA) which overcomes these challenges and produces electronic format from the transformed documents in paper format [2].

Similarly, Optical Character Recognition (OCR) is the process of modification or conversion of any form of text or text-containing documents such as handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Optical character recognition technology enables a machine to automatically recognize text in such documents. In real world example, it is like combination of mind and eye of human body. An eye can detect, view and extract the text from the images but absolutely the human's brain processes that detected or extracted text read by eye [1]. Of course OCR technology has not advanced enough to compete with human's ability. The performance and accuracy of OCR is directly dependent upon the quality of input documents. Again, when we think of human's ability to recognize text, the performance of brain's process directly depends upon the quality of the input read by eye. While designing and implementing a computerized OCR system, several problems and challenges can occur. For example there is very slight difference between some digits and letters for computers to recognize them and distinguish one from the others correctly. For example, it may not be very easy for computers to differentiate between digit "0" and letter "o", especially when these characters are embedded in a very dark and noisy background. One of the main focuses of OCR research has been to recognize cursive scripts and handwritten text for its broad application area. Today, to solve the text recognition problem several different types of OCR software exist such as Desktop OCR, Server OCR, web OCR and so on.

Since the OCR research is an active and important field in general pattern recognition problems, due to its fast progress, comprehensive reviews of the field are needed on a regular basis to keep track of the new advancements. One such review was published to discuss the challenges with text recognition in scene imagery [2]. This paper attempts to elaborate on these kinds of studies by providing a comprehensive literature review of optical character recognition research. We discuss major challenges and main phases of optical character recognition such us pre-processing, segmentation, normalization, feature extraction, classification and post processing in detail which needs to be considered during implementing any application related to the OCR, and in the last section of our paper some OCR applications

[1] *Software Engineering Department, Firat University, Elazig, Turkey*
** Corresponding Author: Email: karez.abdulwahab@gmail.com*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

and a brief OCR history are discussed.

2. OCR Challenges

For good quality and high accuracy character recognition, OCR techniques expect high quality or high resolution images with some basic structural properties such as high differentiating text and background. The way images are generated is an important and determining factor in the accuracy and success of OCR, since this often affects the quality of images dramatically. Usually OCR with images produced by scanners gives high accuracy and good performance. In contrast, images produced by cameras usually are not as good of an input as scanned images to be used for OCR due to the environmental or camera related factors. Numerous errors might emerge, which are clarified as follow.

2.1. Scene Complexity

In a regular environment, we can see large numbers of man-made objects which are included in camera taken images such as paintings, buildings, and symbols. These objects have comparative structures and appearances to text which makes text recognition very challenging in the processed image. Text itself is regularly laid out to encourage decipherability. The challenge with scene intricacy is that the surrounding scene makes it hard to segregate text from non-text [2].

2.2. Conditions of Uneven Lighting

Oftentimes, taking images in natural environments results in uneven lighting and shadows. This poses a challenge for OCR as it degrades the desired characteristics of the image and hence causes less accurate detection, segmentation and recognition results [2].

This condition with uneven lighting is what distinguishes a scanned image form one that is produced with a camera. The lack of such disparities in lighting and shadows makes scanned images preferred over camera images for their better characteristics and quality. Although using an on-camera flash may eliminate such problems with uneven lighting, it introduces new challenges.

2.3. Skewness (Rotation)

For optical character recognition systems, the point of view for the input image that taken from camera of hand-held device or other gadgets that used for taken image is not fixed like a scanner input, which skewing of text lines from their unique orientation might be observed. Great degree poor results will be observed when such a skewed image is fed to the OCR classifier. Many techniques available for the purpose of deskew the image documents, such as Projection Profile, RAST algorithm, Hough transform, methods of Fourier transformation, etc.

2.4. Blurring and Degradation

Since working over a variety of distances are intended to numerous digital cameras, an important factor is the digital camera's focusing. For the best accuracy of character recognition and character segmentation, character sharpness is required. At large apertures and short distances, uneven focus can be observed when a small point of view changes. For the most part connected with photography, there are two kinds of obscure which is: out of focus obscure and movement obscure [3]. At the point for catching a moving item, when the shade rate of the camera is not sufficiently high, the sensor gets presented to a continually changing scene. Accordingly, blurring will observed in parts in motion.

2.5. Aspect Ratios

Text has different aspect ratios. Text may be brief such as traffic signs, while other text may be much longer, such as video captions. Location, scale and length of text need to be considered with search procedure to detect text, which introduces high computational complexity.

2.6. Tilting (Perspective Distortion)

Document images obtained by scanners is constantly parallel to the plane of sensor, but this cannot be observed all times for recorded picture obtained by a portable camera, that may not generally be parallel to the form plane. Accordingly, lines of text that distant from the camera seem littler than those that nearer to the camera which seems greater. This situation causes tilted pictures. Observation of a perspective distortion is clear if the recognizer is not perspective intolerant, which causes lower recognition rate and accuracy [4].

Cell phones have an advantage with orientation sensors. They can recognize whether the device is tilted and when twisting happens they can forbid clients to take pictures. Permitting the user to align plane of the form with that of the camera is also provided by this feature. Orientation sensors therefore may assure that produced pictures satisfy a certain degree of evenness.

2.7. Fonts

Italic style and script fonts of characters might overlap each other, making it difficult to perform some of the main OCR processes such as segmentation. Characters of various fonts have large within-class variations and form many pattern sub-spaces, making it difficult to perform accurate recognition when the character class number is large.

2.8. Multilingual Environments

Albeit a large portion of the languages of Latin have many characters, languages for example, Japanese, Chinese and Korean, have a large number of character classes. Connected characters are exist in Arabic languages, that according to context, it changes writing shape. In Hindi syllables represent by combining alphabetic letters into thousands of shapes [5]. In multilingual situations, OCR in scanned documents stays as a primary research issue [5], since OCR in complex symbolism is more troublesome.

2.9. Warping

Content or text on objects of varying geometries can be another challenge for OCR to be recognized when images of such situation captured by hand-held cameras. A few circumstances may emerge with flatbed scanners, wherein the twisted text observed when the content procured on picture, for example the content towards the binding of an extremely thick book.

For convention paper documents, a technique for image dewarping is proposed by Ulges et al. [6]. By expecting the way that content lines are equally separated and parallel to each other, they dewarp pictures.

3. OCR Phases

In this section we describe the main important phases and architecture of optical character recognition. These phases include pre-processing, segmentation, normalization, future extraction, classification and post processing. For designing an effective application related to the OCR, we must be considering the difficulties that may arise in each phase to obtain high character recognition rate.

3.1. Pre-processing Phase

The aim of pre-processing is to eliminate undesired characteristics or noise in an image without missing any significant information. Preprocessing techniques are needed on colour, grey-level or binary document images containing text and/or graphics. Since processing colour images is computationally more expensive, most of the applications in character recognition systems utilize binary or grey images. Preprocessing reduces the inconsistent data and noise. It enhances the image and prepares it for the next phases in OCR phases.

We can enhance the effectiveness and easiness for an image to be

processed in the next phases by converting the image to the suitable format in the preprocessing phase which is the first phase. Therefore, decreasing the noise that causes the reduction in the character recognition rate is the main important issue in preprocessing phase.
Thus, since preprocessing controls the suitability of the input for the successive phases, a primary stage prior to feature extraction phase is the preprocessing phase. Most of the challenges we listed in OCR Challenges' section need to be addressed in preprocessing stage. Some operations that we may consider to carry out can be listed as follows: binarization, noise reduction, skew correction, morphological operations, slant removal, filtering, thresholding, smoothing, compression, and thinning. Some important preprocessing issues with short description were illustrated in Table 3.1.

Table 1. Some important pre-processing operations

Processes	Description
Binarization	Separates image pixels as text or background.
Noise Reduction	Better improvements of image acquisition devices produced by the advancements in technology.
Skew Correction	Because of the possibility of rotation of the input image through captured image device, document skew should be corrected.
Morphological Operations	Adding or removing pixels to the characters that have holes or surplus pixels.
Thresholding	For an image, separating information from its background.
Thinning and Skeletonisation	thinning process is the Skeletonisation, which regularize the map of the text until reaches most medial one pixel width

3.2. Segmentation Phase

The critical and major component of an Optical Character Recognition (OCR) system is the segmentation of text line from images. In general, Text segmentation from a document image merges line segmentation, word segmentation and then character segmentation. Segmentation is the process of isolating text component within an image from the image's background. For appropriate reorganization of the editable text lines from the recognized characters, firstly, segmenting the line of text, then the words are segmented from the segmented line and then from that the characters are segmented.
Document segmentation is a major pre-processing phase in implementing an OCR system. It is the process of classifying a document image into homogeneous zones, i.e., that each zone contains only one kind of information, such as text, a figure, a table, or a halftone image. In many cases, the accuracy rate of systems related to the OCR heavily depends on the accuracy of the page segmentation algorithm used.
There are three categories of Algorithms of document segmentation [7] As follows:

- Top-down methods,
- Bottom-up methods,
- Hybrid methods.

The top-down approach in a document segments large regions into smaller sub regions recursively. When criterion is met then the document segmentation process will stop and at that stage the ranges obtained constitute the results of final segmentation. But, approaches of bottom-up start by searching for interest pixels and then groups interest pixels. They then manage those interest pixels into connected components that constitute characters which are then combined into words, and lines or text blocks. The integration of both top-down and bottom-up methods is called hybrid approaches.
Regarding different aspects of OCR system throughout the last decades many approaches have already been proposed for segmentation.
A novel technique for Text Segmentation based on A Hough Transform, was proposed by Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu [8].
Extracting Text line from multi skewed handwritten document images has been studied by S. Basu et al [9]. The proposed technique for extraction of text lines implements a water flow technique with high rate of success.
A. Khandelwal et al [10] proposed a technique by comparing neighborhood connected components on text line segmentation from unconstrained handwritten document images.
Shinde, Archana A., and D. G. Chougule also proposed a segmentation technique in their paper [11]. They presented that utilizing the traditional vertical and horizontal projection profile method makes text easily segmented into lines and words. They reported experimental results with 98% accuracy of line and word segmentation.

3.3. Normalization Phase

As a result of segmentation process isolated characters which are ready to move through feature extraction phase are obtained, hence the isolated characters are minimized to a particular size depending on the algorithms used. The segmentation process is crucial as it converts the image in the form of m*n matrix. These matrices are then commonly normalized by minimizing the size and eliminating the unnecessary information from the image without missing any influential information [12].

3.4. Feature Extraction Phase

Feature extraction is the operation of extracting the pertinent features from objects or alphabets to build feature vectors. These feature vectors are then utilized by classifiers to identify the input unit with objective output unit. It becomes effortless for the classifier to classify between dissimilar classes by glancing at these features as it becomes fairly easy to determine [13].
Several techniques are proposed for extracting features from the segmented characters in literature. U. Pal et al [14] have proposed directional chain code features and zoning and for handwritten numeral recognition considered a feature vector of length 100 and have presented a high level of recognition accuracy. But, the feature extraction process is time consuming and complex [13].
Dinesh et al [15] have proposed end points as the potential features for recognition and used horizontal/vertical strokes and for handwritten Kannada numerals obtained a recognition accuracy of 90.50%. But, this method uses the thinning process which results in some loss of features [13].
E. Srinivasan et al [13] for handwritten alphabets recognition system have proposed diagonal based feature extraction utilizing neural network.
Sharma, Om Prakash et al [16] for handwritten alphabets recognition proposed an improved zone based hybrid feature extraction model utilizing Euler number.
Following to Suen [17], there are two major classes of features: statistical features and structural features. In a character matrix statistical features are obtained from statistical distribution of every point such as zoning, moments, crossings, fourier transforms and projection histograms [18]. Statistical features are also notable as global features as they are usually averaged and extracted in sub-images such as meshes. Initially, statistical features are supplied to recognize machine printed characters. On the other hand, structural or topological features are concern to the geometry of the character set to be contemplated. Some of these features are convexities and concavities in the characters, number of holes in the characters, number of end points etc.

3.5. Classification Phase

OCR systems broadly utilize the methodologies of pattern recognition, which assigns each example to a predefined class. Classification is the procedure of distributing inputs with respect to detected information to their comparing class in order to create groups with homogeneous qualities, while segregating different

inputs into different classes. Classification is conveyed out on the premise of put away features in the feature space, for example, structural features, global features and so forth. It can be said that classification isolates the feature space into several classes taking into account the decision rule. Choosing classifier depends on several agents, such as, number of free parameters, available training set and so forth. Various procedures for OCR are explored by the scientists.

Techniques of OCR classification can be categorized [19] as Statistical Techniques, Neural Networks, Template Matching, Support Vector Machine (SVM) algorithms, and Combination of classifier.

3.5.1. Template matching

This is the least complex method for character recognition, in view of matching the stored models against the word or character to be perceived. By gathering of shapes, pixels, curvature and so forth, the operation of matching decides the level of similitude between two vectors. A gray-level or binary input character is contrasted with a standard arrangement of stored models. The recognition rate of this strategy is extremely delicate to noise and input disfigurement.

3.5.2. Statistical Techniques

Hypothesis of Statistical decision is treating with statistical decision capacities and an arrangement of optimality criteria, which for a given model of a specific class can amplify the likelihood of the observed pattern. The main statistical methods that are performed in the area of OCR [19] are Nearest Neighbor (NN), Likelihood or Bayes classifier, Clustering Analysis, Hidden Markov Modelling (HMM), Fuzzy Set Reasoning, and Quadratic classifier.

3.5.3. Neural Networks

Character classification issue is identified with heuristic rationale as people can perceive characters and records by their learning and experience. Thus neural networks which are pretty much heuristic in nature are greatly appropriate for this type of issue [19]. A neural network is an ascertaining architecture that includes enormously parallel interconnection of flexible node processors. Output from one node is reinforcing to the next one in the network and an official choice relies on the complicated collaboration of all nodes. As a result of its similar character, it can apply calculations at a rate higher contrasted with the traditional strategies. Feed-forward neural networks and feedback neural networks can be thought as categorization of neural network architectures. Table 3.2, compares and discusses some recent proposed OCR applications based on Neural Network.

Table 2. Accuracy comparison among recent proposed OCR systems based on neural network

Author(s)	OCR Application	Accuracy %
Shah, Parul, et al [20]	chassis-number recognition	95.49
Zhai, Xiaojun, et al [21]	Automatic Number Plate Recognition ANPR	97.3
Shamsher, Inam, et al [22]	OCR for printed Urdu script	98.3
Yetirajam, Manas, et al [23]	classification and Recognition of broken characters	68.33

3.5.4. Kernel Methods

While the most imperative kernel strategies are support Vector Machines, techniques such as Kernel Fisher Discriminant Analysis (KFDA) and Kernel Principal Component Analysis (KPCA) also employ kernel method. Support vector machines (SVM) are one of the most widely used and most effective supervised learning techniques that can be used for binary or multi-class classification. In classification techniques, by convention the data set first is partitioned into training and testing sets. The objective of SVM is to deliver a model, which predicts the output of the test set [24]. Width of the edge between the classes is the enhancement rule, i.e., the unfilled zone around the decision boundary characterized by the interval to the closest training example [25].

3.5.5. Combination Classifier

Different classification strategies have their own particular advantages and shortcomings. Thus ordinarily various classifiers are consolidated together to solve a given classification problem. Matei, Oliviu, Petrica C. Pop, and H. Vălean [26] by utilizing neural networks and k-Nearest Neighbor, proposed Optical character recognition in real environments such as electricity-meters and gas-meters.

3.6. Postprocessing Phase

It has been shown that people can read handwriting by context up to 60%. While preprocessing tries to clean the record in a specific sense, it might evacuate critical data, since the context data is not accessible at this stage. On the off chance that the semantic data were accessible to a specific degree, it would contribute a considerable measure to the precision of the OCR stages. On the other hand, the whole OCR issue is for deciding the context of the saved image. In this way the incorporation of context and shape data in all the phases of OCR frameworks is vital for meaningful upgrades in recognition rates. This is done in the Postprocessing stage with an input to the early phases of OCR. The least complex method for consolidating the context data is the usage of a dictionary for amending the minor errors of the OCR frameworks. The fundamental thought is to spell check the OCR yield and give a few distinct options for the yields of the recognizer that take place in the dictionary.

4. OCR Applications

Optical character recognition has been performed in a numerous of applications. We discussed some of these application areas in this section.

4.1. Handwriting Recognition

Handwriting recognition is the capacity of a PC to get and translate intelligible handwritten data from sources, for example, paper records, photos, touch-screens and different gadgets. The picture of the written content might be detected "off line" from a bit of paper by optical scanning (optical character recognition) or clever word recognition. On the other hand, the developments of the pen tip may be detected "on line", for instance by a pen-based PC screen surface.

4.2. Receipt Imaging

Receipt imaging [27] is broadly utilized as a part of numerous organizations applications to monitor financial records and keep accumulation of payments from heaping up. In government offices and autonomous organizations, OCR simplifies information gathering and analysis, among different procedures.

4.3. Legal Industry

Legal industry [27] is likewise one of the recipients of the OCR innovation. OCR is utilized to digitize documents, and to specifically enter into PC database. Legitimate experts can further search documents required from tremendous databases by basically writing a few keywords.

4.4. Banking

Another imperative use of OCR is in banking [27], where it is utilized to process cheques without human intervention. A cheque can be embedded with a machine where the framework filters the sum to be issued and the right measure of cash is exchanged. This

innovation has been idealized for printed cheque, and is genuinely precise for handwritten checks diminishing the hold-up time in banks.

4.5. Healthcare

To process printed material, medicinal services [27] have likewise seen an expansion in the utilization of OCR innovation. Medicinal service experts continuously need to manage extensive volumes of documents for each patient, including protection frames and in addition general health forms. To stay aware of every one of this data, it is valuable to input relevant information into an electronic database. With OCR processing tools, we can extract data from structures and put it into databases, so that each patient's information is quickly recorded and retrieved when needed in future.

4.6. Captcha

A CAPTCHA [28] is a system that can create and grade tests that human can pass yet current software technology can't. Malicious programmer can make software to misuse personal information on websites. Dictionary attack is assault against secret word confirmed frameworks where a programmer composes a system to over and over attempt distinctive passwords like from dictionary of most regular passwords. In CAPTCHA, a picture comprising an arrangement of letters and numbers is produced with variety of size and textual styles, distracting backgrounds, arbitrary portions, highlights and noise so that text cannot be read via OCR. Current OCR frameworks can be utilized to evacuate the noise and portion the picture to make the picture tractable by such malicious users.

4.7. Automatic Number Plate Recognition

Automatic number plate recognition [29] is utilized as a mass observation method making utilization of optical character recognition on pictures to recognize vehicle registration plates. ANPR has additionally been made to store the pictures caught by the cameras including the numbers caught from license plate. ANPR innovation own to plate variety from place to place as it is an area particular innovation. They are utilized by different police forces and as a technique for electronic toll accumulation on pay-per-use streets.

4.8. ATMA: android travel mate application

ATMA: android travel mate application [30] proposed by Mishra, Nitin, and C. Patvardhan, that It empowers Tourists and Travellers to effortlessly catch the native signboards, nation dialect Books pages, hotel menus, banners and so on. Unicode text format was obtained from content embedded in the caught image by an implicit OCR. With the goal that Travellers can translate the native Dialect Unicode content into their own nation dialect, it likewise gives translation feature.

5. OCR History

The innovation of retina scanner is the initial character recognition concept, that it is a framework for picture transmission which makes utilization of photocells mosaic [31]. A noteworthy leap forward happened in 1890 which is the today's TV and perusing machines with Nipkow's innovation which is a successive scanner. Amid the OCR early ages was considered as a help for the visually impaired individuals. But, later on it formed into an unfathomable field of innovative work.

in Germany in 1929, a patent recording of Tauschek [32] is the primary proof of optical character recognition framework and later in 1935, he was stated US Patent and in 1933, Handel was stated openly [33]. Template symbols with a circular disc are utilized by both the machines, so that light sparkles can be observed through it. In front of the disc, the picture that needs to be perceived is held and is then lit up. Through the template hole, the light reflecting off a part of it is then engaged and a photo sensor is utilized for distinguishing it.

The industrially accessible OCRs might be arranged into four generations based on their strength, effectiveness and adaptability. Only chosen text styles and shapes of characters could be read by the original OCRs. Such machines were utilized in mid 1960s. IBM 1418 [34] was the first generation OCR to become into generally marketed. Logical template matching was the method that utilized. The OCRs second generations were substantially more fit, which could perceive both characters printed by machine and additionally handwritten. The OCRs second generations which were accessible amid the center of 1960 to mid-1970s was confined to numerals only. IBM 1287 was the main OCR arrangement of second generation, which was a mixture framework that consolidates both analog and digital technology [34].

For improvement of print based scanned document further research was done and consequently built up the OCRs third generation. They work on hand-written characters of poor print quality characters than before and huge set. Such frameworks were well known amid the period 1975-1985 [34].

For filtering and perceiving characters from complex documents intermixed with writings, The OCRs fourth generation are fit, Mathematical symbols and tables furthermore unconstrained handwritten characters, low quality noisy documents, for example, photocopies, fax and color documents. Now more complex OCRs are accessible for Arabic, Chinese, Japanese and Roman primer [31, 35, 36].

6. Conclusion

Numerous algorithms, methods and techniques have been proposed to optical character recognition in scene imagery, yet there are not enough literature surveys in this field. In this paper, we have proposed an organization of these methods, algorithms and techniques. It is hoped that this comprehensive survey will provide insight into the concepts involved, and perhaps provoke further advances in the area. Firstly, we discussed major challenges of OCR, then we discussed in great detail the main important phases, architecture, proposed algorithms and techniques of OCR, we highlight that for designing any application related to the OCR, one must pay great attention to each phase to obtain high accurate character recognition rate, but still we cannot propose comprehensive algorithms for each phase because it depends upon datasets, application specifics, and parameter specifics. Finally major applications related to the OCR and a brief OCR history are discussed.

Although the state-of-the art OCR enables text recognition with high accuracy, we think that there could be many more practical applications of OCR. As a future work we are planning to use OCR for such practical applications for daily personal use. We are planning to incorporate mobile devices with OCR in one OCR system. An automated book reader or a receipt tracker constitutes some of our future OCR based applications.

References

[1] Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications. 2012 Jan 1;55(10).

[2] Ye Q, Doermann D. Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence. 2015 Jul 1;37(7):1480-500.

[3] Jain A, Dubey A, Gupta R, Jain N, Tripathi P. Fundamental challenges to mobile based ocr. vol. 2013 May;2:86-101.

[4] Moravec K. A Grayscale Reader for Camera Images of Xerox DataGlyphs. InBMVC 2002 Sep (pp. 1-10).

[5] Smith R, Antonova D, Lee DS. Adapting the Tesseract open source OCR engine for multilingual OCR. In Proceedings of the International Workshop on Multilingual OCR 2009 Jul 25 (p. 1). ACM.

[6] Ulges A, Lampert CH, Breuel TM. Document image

dewarping using robust estimation of curled text lines. In Eighth International Conference on Document Analysis and Recognition (ICDAR'05) 2005 Aug 31 (pp. 1001-1005). IEEE.
[7] Kaur S, Mann PS, Khurana S. Page Segmentation in OCR System-A Review.
[8] Saha S, Basu S, Nasipuri M, Basu DK. A Hough transform based technique for text segmentation. arXiv preprint arXiv:1002.4048. 2010 Feb 22.
[9] Basu S, Chaudhuri C, Kundu M, Nasipuri M, Basu DK. Text line extraction from multi-skewed handwritten documents. Pattern Recognition. 2007 Jun 30;40(6):1825-39.
[10] Khandelwal A, Choudhury P, Sarkar R, Basu S, Nasipuri M, Das N. Text line segmentation for unconstrained handwritten document images using neighborhood connected component analysis. InInternational Conference on Pattern Recognition and Machine Intelligence 2009 Dec 16 (pp. 369-374). Springer Berlin Heidelberg.
[11] Shinde AA, Chougule DG. Text Pre-processing and Text Segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810-2.
[12] Trier ØD, Jain AK, Taxt T. Feature extraction methods for character recognition-a survey. Pattern recognition. 1996 Apr 30;29(4):641-62.
[13] Pradeep J, Srinivasan E, Himavathi S. Diagonal based feature extraction for handwritten character recognition system using neural network. InElectronics Computer Technology (ICECT), 2011 3rd International Conference on 2011 Apr 8 (Vol. 4, pp. 364-368). IEEE.
[14] Bishnu A, Bhattacharya BB, Kundu MK, Murthy CA, Acharya T. A pipeline architecture for computing the Euler number of a binary image. Journal of Systems Architecture. 2005 Aug 31;51(8):470-87.
[15] Dinesh Acharya U, Subbareddy NV. Krishnamoorthy: Isolated Kannada Numeral Recognition Using Structural Features and K-Means Cluster. Proc. of IISN. 2007:125-9.
[16] Sharma OP, Ghose MK, Shah KB. An improved zone based hybrid feature extraction model for handwritten alphabets recognition using euler number. International Journal of Soft Computing and Engineering. 2012 May;2(2):504-8.
[17] Suen CY. Character recognition by computer and applications. Handbook of pattern recognition and image processing. 1986:569-86.
[18] Rehman A, Saba T. Neural networks for document image preprocessing: state of the art. Artificial Intelligence Review. 2014 Aug 1;42(2):253-73.
[19] Dongre VJ, Mankar VH. A review of research on Devnagari character recognition. arXiv preprint arXiv:1101.2491. 2011 Jan 13.
[20] Shah P, Karamchandani S, Nadkar T, Gulechha N, Koli K, Lad K. OCR-based chassis-number recognition using artificial neural networks. InVehicular Electronics and Safety (ICVES), 2009 IEEE International Conference on 2009 Nov 11 (pp. 31-34). IEEE.
[21] Zhai X, Bensaali F, Sotudeh R. OCR-based neural network for ANPR. In2012 IEEE International Conference on Imaging Systems and Techniques Proceedings 2012 Jul 16 (pp. 393-397). IEEE.
[22] Shamsher I, Ahmad Z, Orakzai JK, Adnan A. OCR for printed urdu script using feed forward neural network. InProceedings of World Academy of Science, Engineering and Technology 2007 Aug (Vol. 23, pp. 172-175).
[23] Yetirajam M, Nayak MR, Chattopadhyay S. Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume. 2012 Oct 28;1.
[24] Verma R, Ali DJ. A-Survey of Feature Extraction and Classification Techniques in OCR Systems. International Journal of Computer Applications & Information Technology. 2012 Nov;1(3).
[25] Jain AK, Duin RP, Mao J. Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence. 2000 Jan;22(1):4-37.
[26] Matei O, Pop PC, Vălean H. Optical character recognition in real environments using neural networks and k-nearest neighbor. Applied intelligence. 2013 Dec 1;39(4):739-48.
[27] Ganis MD, Wilson CL, Blue JL. Neural network-based systems for handprint OCR applications. IEEE Transactions on Image Processing. 1998 Aug;7(8):1097-112.
[28] Gossweiler R, Kamvar M, Baluja S. What's up CAPTCHA?: a CAPTCHA based on image orientation. In Proceedings of the 18th international conference on World wide web 2009 Apr 20 (pp. 841-850). ACM.
[29] Gao J, Blasch E, Pham K, Chen G, Shen D, Wang Z. Automatic vehicle license plate recognition with color component texture detection and template matching. In SPIE Defense, Security, and Sensing 2013 May 21 (pp. 87390Z-87390Z). International Society for Optics and Photonics.
[30] Mishra N, Patvardhan C. ATMA: Android Travel Mate Application. International Journal of Computer Applications. 2012 Jan 1;50(16).
[31] Mantas J. An overview of character recognition methodologies. Pattern recognition. 1986 Dec 31;19(6):425-30.
[32] Gustav Tauschek. Reading machine. U.S. Patent 2026329, http://www.google.com/patents?vid=USPAT2026329, December 1935FLEXChip Signal Processor (MC68175/D), Motorola, 1996. [Accessed 23/11/2016]
[33] Paul W. Handel. Stat IST ical Machine. U.S. Patent 1915993,http://www.google.com/patents?vid=USPAT1915993, June 1993. [Accessed 23/11/2016]
[34] Mori S, Suen CY, Yamamoto K. Historical review of OCR research and development. Proceedings of the IEEE. 1992 Jul;80(7):1029-58.
[35] Amin A. Off-line Arabic character recognition: the state of the art. Pattern recognition. 1998 Mar 1;31(5):517-30.
[36] Stallings W. Approaches to Chinese character recognition. Pattern recognition. 1976 Apr 30;8(2):87-98.