A SURVEY ON OPTICAL CHARACTER RECOGNITION SYSTEM

Osamah Adil Raheem Almusawi

University of Wasit/ College of Computer Science and Information Technology

oalmusawi@uowasit.edu.iq

Abstract

Optical Character Recognition (OCR) is a piece of software that converts printed text and images into digitized form such that it can be manipulated by machine. Unlike human brain which has the capability to very easily recognize the text/ characters from an image, machines are not intelligent enough to perceive the information available in image. Methodology this paper discusses different types of OCR systems have emerged as a result of multitude of directions in which research on OCR has been carried out during past years. In this paper, an overview of various techniques of OCR has been presented. An OCR is not an atomic process but comprises various phases such as acquisition, pre- processing, segmentation, feature extraction, classification and post-processing. Each of the steps is discussed in detail in this paper. Using a combination of these techniques, an efficient OCR system can be developed as a future work. The OCR system can also be used in different practical applications such as number-plate recognition, smart libraries and various other real-time applications. Despite of the significant amount of research in OCR, recognition of characters for language such as Arabic, Sindhi and Urdu still remains an open challenge

Introduction

Optical Character Recognition (OCR) is a piece of software that converts printed text and images into digitized form such that it can be manipulated by

machine. Unlike human brain which has the capability to very easily recognize the text/ characters from an image, machines are not intelligent enough to perceive the information available in image. Therefore, a large number of research efforts have been put forward that attempts to transform a document image to format understandable for machine. OCR is a complex problem because of the variety of languages, fonts and styles in which text can be written, and the complex rules of languages etc. Hence, techniques from different disciplines of computer science (i.e. image processing, pattern classification and natural language processing etc. are employed to address different challenges. This paper introduces the reader to the problem. It enlightens the reader with the historical perspectives, applications, challenges and techniques of OCR.

Literature review

Character recognition is not a new problem but its roots can be traced back to systems before the inventions of computers. The earliest OCR systems were not computers but mechanical devices that were able to recognize characters, but very slow speed and low accuracy. In 1951, M. Sheppard invented a reading and robot GISMO that can be considered as the earliest work on modern OCR. GISMO can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the capability to could copy a typewritten page. J. Rainbow, in 1954, devised a machine that can read uppercase typewritten English characters, one per minute. The early OCR systems were criticized due to errors and slow recognition speed. Hence, not much research efforts were put on the topic during 60's and 70's. The only developments were done on government agencies and large corporations like banks, newspapers and airlines etc. Because of the complexities associated with recognition, it was felt that three should be standardized OCR fonts for easing the task of recognition for OCR. Hence, OCRA and OCRB were developed by ANSI and EMCA in 1970, that provided comparatively acceptable recognition rates. During the past thirty

years, substantial research has been done on OCR. This has led to the emergence of document image analysis (DIA), multi-lingual, handwritten and omni-font OCRs. Despite these extensive research efforts, the machine's ability to reliably read text is still far below the human. Hence, current OCR research is being done on improving accuracy and speed of OCR for diverse style documents printed/written in unconstrained environments. There has not been availability of any open source or commercial software available for complex languages like Urdu or Sindhi etc.

Methodology

There has been multitude of directions in which research on OCR has been carried out during past years. This section discusses different types of OCR systems have emerged as a result of these researches. We can categorize these systems based on image acquisition mode, character connectivity, font-restrictions etc. Fig. 1 categorizes the character recognition system. Based on the type of input, the OCR systems can be categorized as handwriting recognition and machine printed character recognition. The former is relatively simpler problem because characters are usually of uniform dimensions, and the positions of characters on the page can be predicted. Handwriting character recognition is a very tough job due to different writing style of user as well as different pen movements by the user for the same character. These systems can be divided into two sub-categories i.e. on-line and off-line systems. The former is performed in real-time while the users are writing the character. They are less complex as they can capture the temporal or time based information i.e. speed, velocity, number of strokes made, direction of writing of strokes etc. In addition, there no need for thinning techniques as the trace of the pen is few pixels wide. The offline recognition systems operate on static data i.e. the input is a bitmap. Hence, it is very difficult to perform recognition. There have been many online systems available because they are

easier to develop, have good accuracy and can be incorporated for inputs in tablets and PDAs.
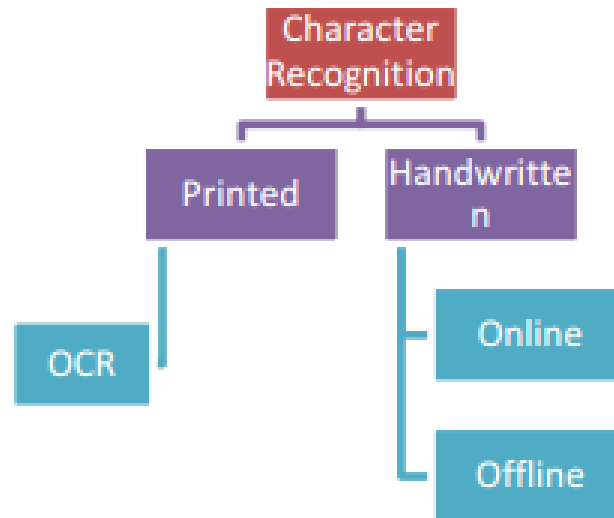

Figure.1: Types of character recognition system IV.

Applications of OCR

OCR enables a large number of useful applications. During the early days, OCR has been used for mail sorting, bank cheque reading and signature verification. Besides, OCR can be used by organizations for automated form processing in places where a huge number of data is available in printed form. Other uses of OCR include processing utility bills, passport validation, pen computing and automated number plate recognition etc. Another useful application of OCR is helping blind and visually impaired people to read text.

Major phases of OCR

The process of OCR is a composite activity comprises different phases. These phases are as follows: Image acquisition: To capture the image from an external source like scanner or a camera etc. Preprocessing: Once the image has been acquired, different preprocessing steps can be performed to improve the quality of image. Among the different preprocessing techniques are noise removal, thresholding and extraction image base line etc. Character segmentation: In this step, the characters in the image are separated such that they can be passed to

recognition engine. Among the simplest techniques are connected component analysis and projection profiles can be used. However, in complex situations, where the characters are overlapping /broken or some noise is present in the image. In these situations, advance character segmentation techniques are used. Feature extraction: The segmented characters are then processes to extract different features. Based on these features, the characters are recognized. Different types of features that can be used extracted from images are moments etc. The extracted features should be efficiently computable, minimize intra-class variations and maximizes inter-class variations. Character classification: This step maps the features of segmented image to different categories or classes. There are different types of character classification techniques. Structural classification techniques are based on features extracted from the structure of image and uses different decision rules to classify characters. Statistical pattern classification methods are based on probabilistic models and other statistical methods to classify the characters. Post processing: After classification, the results are not 100% correct, especially for complex languages. Post processing techniques can be performed to improve the accuracy of OCR systems. These techniques utilizes natural language processing, geometric and linguistic context to correct errors in OCR results. For example, post processor can employ a spell checker and dictionary, probabilistic models like Markov chains and n-grams to improve the accuracy. The time and space complexity of a post processor should not be very high and the application of a post-processor should not engender new errors.

1) Image Acquisition

Image acquisition is the initial step of OCR that comprises obtaining a digital image and converting it into suitable form that can be easily processed by computer. This can involve quantization as well as compression of image. A special case of quantization is binarization that involves only two levels of image. In most of the cases, the binary image suffices to characterize the image. The

compression itself can be loss or loss-less. An overview of various image compression techniques have been provided in.

2) Pre-processing

Next to image acquisition is pre-processing that aims to enhance the quality of image. One of the pre-processing techniques is thresholding that aims to binaries the image based on some threshold value. The threshold value can be set at local or global level. Different types of filters such as averaging, min and max filters can be applied. Alternatively, different morphological operations such as erosion, dilation, opening and closing can be performed.

An important part of pre-processing is to find out the skew in the document. Different techniques for skew estimation includes: projection profiles, Hough transform, nearest neighborhood methods. In some cases, thinning of the image is also performed before later phases are applied. Finally, the text lines present in the document can also be found out as part of pre-processing phase. This can be done based on projections or clustering of the pixels.

3) Character Segmentation

In this step, the image is segmented into characters before being passed to classification phase. The segmentation can be performed explicitly or implicitly as a byproduct of classification phase. In addition, the other phases of OCR can help in providing contextual information useful for segmentation of image.

4) Feature Extraction

In this stage, various features of characters are extracted. These features uniquely identify characters. The selection of the right features and the total number of features to be used is an important research question. Different types of features such as the image itself, geometrical features (loops, strokes) and statistical feature (moments) can be used. Finally, various techniques such as

principal component analysis can be used to reduce the dimensionality of the image.

5) Classification

It is defined as the process of classifying a character into its appropriate category. The structural approach to classification is based on relationships present in image components. The statistical approaches are based on use of a discriminate function to classify the image. Some of the statistical classification approaches are Bayesian classifier, decision tree classifier, neural network classifier, nearest neighborhood classifiers etc. Finally, there are classifiers based on syntactic approach that assumes a grammatical approach to compose an image from its sub-constituents.

6) Post-processing

Once the character has been classified, there are various approaches that can be used to improve the accuracy of OCR results. One of the approaches is to use more than one classifier for classification of image. The classifier can be used in cascading, parallel or hierarchical fashion. The results of the classifiers can then be combined using various approaches. In order to improve OCR results, contextual analysis can also be performed. The geometrical and document context of the image can help in reducing the chances of errors. Lexical processing based on Markov models and dictionary can also help in improving the results of OCR.

Phase	Description	Approaches
Acquisition	The process of acquiring image	Digitization, binarization, compression
Pre-processing	To enhance quality of image	Noise removal, Skew removal, thinning, morphological operations
Segmentation	To separate image into its constituent characters	Implicit Vs Explicit Segmentation
Feature Extraction	To extract features from image	Geometrical feature such as loops, corner points Statistical features such as moments
Classification	To categorize a character into its particular class	Neural Network, Bayesian, Nearest Neighborhood
Post- processing	To improve accuracy of OCR results	Contextual approaches, multiple classifiers, dictionary based approaches

Conclusions

In this paper, an overview of various techniques of OCR has been presented. An OCR is not an atomic process but comprises various phases such as acquisition, pre- processing, segmentation, feature extraction, classification and post-processing. Each of the steps is discussed in detail in this paper. Using a combination of these techniques, an efficient OCR system can be developed as a future work. The OCR system can also be used in different practical applications such as number-plate recognition, smart libraries and various other real-time

applications. Despite of the significant amount of research in OCR, recognition of characters for language such as Arabic, Sindhi and Urdu still remains an open challenge. An overview of OCR techniques for these languages has been planned as a future work. Another important area of research is multi-lingual character recognition system. Finally, the employment of OCR systems in practical applications remains an active are of research.

References

1) Cap Digital. Intelligence Artificielle: technologies, marches et defis.

2) Document # 15-167, www.capdigital.com/publications, 2015

3) Ernest Davis and Gary Marcus. Commonsense Reasoning and Commonsense

4) Knowledge in Artificial Intelligence. Communications Of The ACM Vol. 58 No. 9, 2015

5) Jonathan Grudin. AI and HCI: Two Fields Divided by a Common Focus. AI magazine, 30(4), 48-57, 2008

6) Kevin Kelly. The Three Breakthroughs That Have Finally Unleashed AI On The World. http://www.wired.com/2014/10/future-of-artificial-intelligence/, 2014

7) Pierre Marquis, Odile Papini, Henri Prade (eds). Panorama de l'Intelligence

8) Artificielle. ses bases méthodologiques, ses développements. 3 vols. Cepaduès, 2014

9) Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. http://aima.cs.berkeley.edu

10) Alan Turing. Intelligent Machinery, a Heretical Theory. Philosophia Mathematica 4 (3): 256-260, 1996 - Original article from 1951

11) Terry Winograd. Shifting viewpoints: Artificial intelligence and human–computer interaction. Artificial Intelligence 170(18):1256-1258, 2006

12) Ronald C. Arkin. The Case for Ethical Autonomy in Unmanned Systems. Journal of Military Ethics, 9(4), 12/2010

13) Samuel Butler. Erewhon. Free eBooks at Planet eBook.com, 1872

14) Dominique Cardon. A quoi rêvent les algorithmes. Seuil, 2015

15) Thomas G. Dietterich and Eric J. Horvitz. Rise of Concerns about AI: Reflections and Directions. Communications of the ACM | Vol. 58 No. 1, October 2015

16) Moshe Vardi. On Lethal Autonomous Weapons. Communications of the ACM, vol. 58 no. 12, December 2015

17) Martin Abadi et al. Large-Scale machine learning on Heterogeneous Distributed

18) Systems. Software available from tensorflow.org, 2015

19) Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006

20) Léon Bottou. From machine learning to machine reasoning: an essay, Machine Learning, 94:133-149, January 2014

21) Yann Le Cun. The Unreasonable Effectiveness of Deep Learning. Facebook AI Research & Center for Data Science, NYU. http://yann.lecun.com, 2015

22) Sumit Gulwani, William R. Harris, and Rishabh Singh. Spreadsheet Data Manipulation Using Examples. Communications of the ACM, Vol. 55 no. 8, 2012

23) Michael I. Jordan and Tom M. Mitchell. Machine learning: trends, perspectives, and prospects. Science, Vol. 349 Issue 6245, 2015

24) James Max Kanter and Kalyan Veeramachaneni. Deep Feature Synthesis: Towards Automating Data Science Endeavors. International IEEE/ACM Data Science and Advance Analytics Conference, 2015

25) Tom M. Mitchell. The Discipline of Machine Learning. CMU-ML-06-108 School of Computer Science Carnegie Mellon University, 2006

26) Volodymyr Mnih et al. Human-level control through deep reinforcement learning. Nature 518, 529–533, 2015

27) Michèle Sebag. A tour of Machine Learning: an AI perspective.

28) Peter Auer et al. A Research Roadmap of Cognitive Vision.

29) ECVision : the European Research network for Cognitive Computer Vision Systems. https://www.researchgate.net, 2005