Handwriting/Digit Recognition Using OpenCV Python

Kunal Jagdale,Tanmay Jagtap, Vijay Gaikwad, Milind Rane

kunal.jagdale15@vit.edu, tanmay.jagtap15@vit.edu, vijay.gaikwad@vit.edu, milind.rane@vit.edu

Department of Electronics, Vishwakarma Institute of Technology, Pune

ABSTRACT- ***Computer vision and machine learning are two domains that are upcoming in the modern era. Computer vision is designed to try to make computer as good as human. There are many things we humanshave in common. But there are other things that are unique to every individual - DNA, fingerprints, etc. Handwriting is something that is different to every individual, which is already proven in the recent studies on Handwriting analysis. Although arguable is this issue, that handwriting can be mimicked and forgery becoming a huge issue, there is certain level of individuality and uniqueness that cannot be mimicked or forged. As computerization is becoming more prominent these days, Handwriting Recognition is gaining importance in various fields e.g.Authentication of signatures in banks, recognizing ZIP codes addresses on letters, forensic evidence, etc. Furthermore, letting a large-scale computational system do all the analysis and the authentication work in the bank and other agencies reduced much of the burden. But how can a computer, which is a machine recognize the handwriting of an individual? Becauseeach individual has his own way of presenting his/her ideas on paper, there is a certain level of complexity involved in this subject. This paper presents an overview of some methodologies and recognition algorithms, particularly off-line recognition methods. Optical character recognition is a mechanism in which the handwritten text can be converted to machine readable data. OCR has gained a lot of importance for this purpose of handwriting recognition. Support vector machine (SVM), is used for recognizing different types of patterns. SVM is well known for recognizing different patterns successfully. MINST digit database has been used, which is huge database of handwritten digits of different types of handwritings.***

Keywords- Handwriting identification, feature extraction, handwriting individuality, OCR, MINST database

1. INTRODUCTION

All the modern inventions in computer and communication technologies such as word processors, fax machines and e-mail are having their impact on handwriting [1]. These changes have led to the tuning and reinterpreting of the role of handwriting and handwrittenmessages.

Despite these modern marvels, a pen together with a paper is much more convenient than a keyboard or a mouse. Computers which will process handwriting must have to deal with different writing styles and handwritings, work with random alphabets which are user-defined, and understand any handwritten message by any writer [2].

Handwriting recognition is an ability and technique of the system which takes input from electronic pen, touch screen,images, scanner and paper documents. Offline handwriting recognition system is an art of identifying the digits and numbers from images. Handwriting recognition system is developed for accuracy and reliability. So handwriting recognition is one of the most challenging areas in image and pattern recognition. Handwriting recognition is very useful in real world. There are many real-life problems where handwriting recognition system is very useful like mailing address interpretation, signature verification, documentation analysis, bank check processing, postal addresses. There are a number of approaches which have been used both in online and offline handwriting recognition field like statistical methods, structural methods, neural network and syntactic methods. Some recognition systemapply recognition, other identify strokes on single character or entire words.

Handwriting data gets converted into digital form either by writing with a special pen on an electronic surface or by scanning the writing on paper. The two approaches are distinguished as off- line and on-line handwriting, respectively. In the on-line case, the two-dimensional co- ordinates of successive points of the writing as a function of time are stored in order[3]. In case of the offline, only the completed writing is available as an image. The recognition rates reported are much higher for the on-line case in comparison with the off-line case. Off-line systems are less accurate than on-line systems. However, they are now good enough that they have a significant economic impact for specialized domains such as interpreting hand-written postal addresses on envelopes and reading courtesy amounts on bank checks.

The success of on-line systems makes it attractive to consider developing, off-line systems that first estimate the trajectory of the writing from off- line data and then use online algorithms.

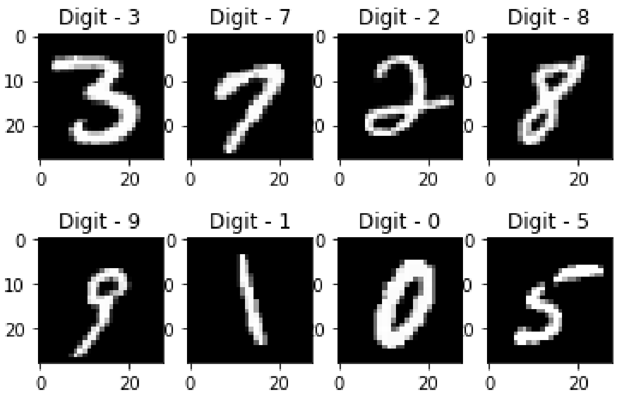

Fig.1: Training data

The field of off-line handwritten word and digit recognition has advanced greatly in the last 10 years and thus is the core focus of this paper. Many different approaches have been proposed and implemented by researchers. In the literature, performance of the handwritten word / digit recognizers is generally reported as accuracy rates on lexicons of different sizes, eg., 10, 100 and 1000 [3].

MINST: The MINST database is a big database of handwritten characters and digits. This is commonly used for training image processing systems. MINST is now standard training dataset for training in the English language. There are similar databases by others for training in different languages.

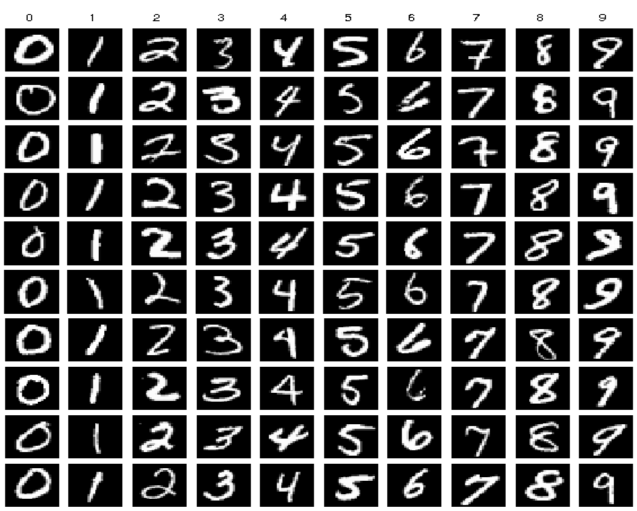

Fig 2. Sample database of MINST

2. Offline Handwriting DigitRecognition

The primary functions of offline handwriting recognition are character recognition and word recognition. The necessary preliminary step in recognition is document analysis which locates appropriate text when complex, two-dimensional spatial lay-outs are employed [1]. Different approaches have been proposed to offline recognition that have contributed to the present day efficiency of thetechnique.

2.1 Preprocessing

Several document analysis operations must be performed prior to recognizing text in scanned documents. Some of the common operations performed prior to recognition are: thresholding,the ex-traction of the foreground textual matter by removing, say, textured background, the task of converting a gray-scale image into abinary black-white image; noise removal, salt and pepper noise and interfering strokes; line segmentation, the separation of individual lines of text; word segmentation, the isolation of textual words, the isolation of individual character,and character segmentation, typically those words which are written discretely rather thancursively.

2.1.1 Thresholding

The task of thresholding is to extract the foreground (ink) from the background (paper). The histogram of gray-scale values of a document image typically consists of two peaks: a high peak corresponding to the white background and a smaller peak corresponding to the fore-ground. So, the task of determining the threshold gray-scale value is one of determining an “optimal” value in the valley between the two peaks[1].

The distributions of the background andforeground points are classified as two classes. Each value of the threshold is tried and one that maximizes the criterion is chosen. There are several improvements to this basic idea, such as handling textured backgrounds similar to those encountered on bank checks.

2.1.2 NoiseRemoval

Noise removal is a prerequisite in document analysis which has been dealt with extensively for typed or even machine-printed documents. For handwritten documents, the connectivity of strokes has to be preserved. The images which are digitally captured can introduce noise from scanning devices and transmission media. One study, which is given below,provides a method which performs selective and adaptive stroke “filling” with a neighborhood operator. This emphasizes stroke connectivity, while also, conservatively checking aggressive“over-filling.”[1]

2.1.3 LineSegmentation

Segmentation of handwritten text into lines, digits, and characters has many sophisticated approaches. This is completely opposite to the task of segmenting lines of text into words and characters, which is easier for machine-printed documents. It can be accomplished by examining the horizontal histogram profile at a small range of skew angles. The task is more difficult in the handwritten domain. Here, lines of text might be undulate up and down and ascenders and descenders frequently intersect characters of neigh-boring lines. One of the method is based on the principle that people write on an imaginary line which forms the base upon which each word of the line resides. The local minima points approximate this imaginary baseline from each component. A clustering technique is used to group the minima of all the components to identify the different handwrittenlines.[1]

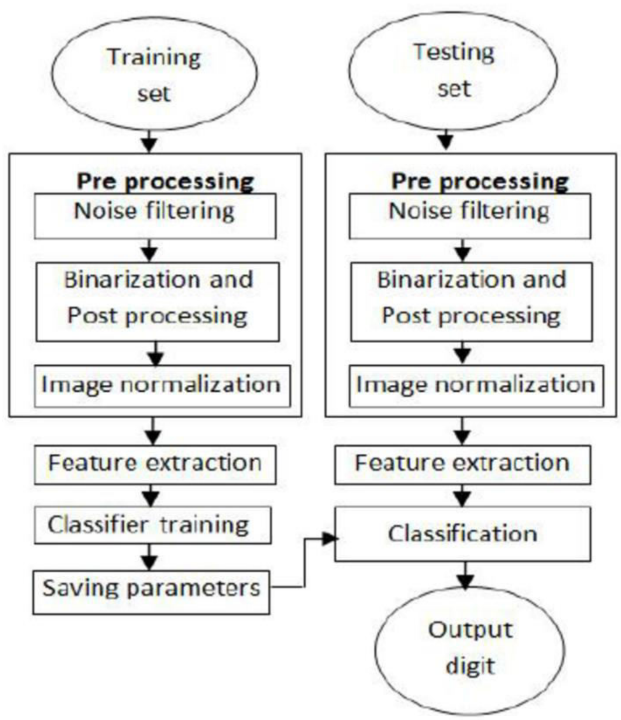

Fig. 3: Digit Recognition Algorithm

2.2 Digit Character Recognition

One of the most basic and common problem is to assign the digitized character to its own symbolic class. In case of a print image, this is known to as Optical Character Recognition (OCR) [1]. In the case of handprint, it is loosely referred to as intelligent character recognition (ICR) [1]. We limit our research to the recognition of English orthography in the handwritten form.

Most character recognition techniques use this approach, a set of features and a classification method are developed and every test pattern is subject to the same process, called asa "one model fits all" approach, whichever of the constraints present in the problem domain[2].

For extracting shape features and to assign the observed character to the appropriate class, a pattern recognition algorithm is used to. Artificial neural networks have emerged as fast methods for implementing classifiers for OCR[1],[3]. Character recognition from a single, machine-printed font family on a well-printed paper document can be done very accurately. Difficulties arise when handwritten characters are to be handled. In difficult cases, it becomes necessary to use models to constrain the choices at the character and word levels. These type of models are very useful in handwriting recognition due to the wide variety of hand printing and cursivescript.

Given a handwriting sample, a set of digits is first segmented, then for each isolated character, the so-called micro-features are extracted. Therefore, each and every handwriting sample is characterized by a number of micro-feature vectors corresponding to the characters available from the sample.

Feature extraction methods are the core to achieve high-performing word recognition. One approach utilizes the idea of both "singular" and "regular" features. Any form of handwriting is regarded to have a regular flow modified by occasional singular embellishments. A widely used approach is to use an HMM to structure the entire recognitionprocess.

Digits that are relevant during the recognition task are not available during training because they belong to an unknown subset of a very large lexicon. Individual characters are over segmented such that after the segmentation process no adjacent characters remain touching. Instead of passing on combi- nations of segments to a generic OCR, a lexicon is brought into play early in the process. A combination of adjacent segments is compare to only those character choices which are possible at the position in the word being considered. The approach can be viewed as a process of accounting for all the segments generated by a given lexicon entry. Lexicon entries are ordered according to the "goodness" of the

International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 04 | April -2019 ISSN: 1847-9790 || p-ISSN: 2395-0126

match [1],[5].

A commonly and very widely used paradigm to string the potential character candidates into word candidates is Dynamic Programming (DP); some methods combine heuristics with DP to disqualify certain groupsofprimitivesegmentswhich areevaluatediftheyaretoo complex to represent a single character. The compatibility between consecutive character candidates is also taken into account by the DP paradigm [1].

3. Results

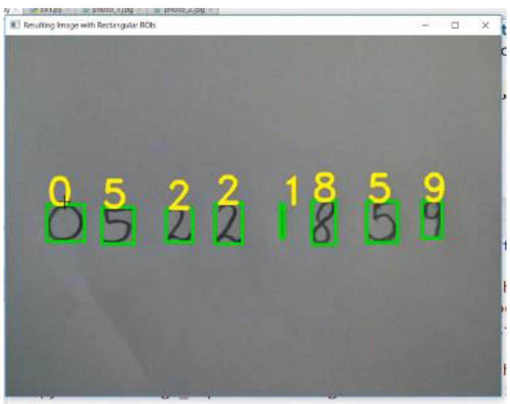

Fig 4. Handwritten digits recognized

4. Conclusion

Research on automated written language recognition dates back many years to even several decades. Presently, cleanly machine-printed text documents with simple layouts can be recognized accurately by off-the-shelf OCR software [1], [3]. There is also some success with handwriting recognition, particularly for isolated hand printed characters and words, which we have seen throughout this paper. For example, in the online case, the recently introduced PDAs have got some practical value. Similarly, some on-line signature verification systems have been marketed over the last few years and instructional tools to help children learn to write are beginning to emerge.

In an e-world dominated by the WWW, the design of human- computer interfaces based on handwriting is part of a tremendous research effort together with speech recognition, language processing and translation to facilitate communication of people with computer networks. From this perspective, any successes or failure in these fields will have a great impact on the evolution of languages.

5. References

[1] Plamondon, Réjean, and Sargur N. Srihari. "Online and off-line handwriting recognition: a comprehensive survey." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.1 (2000):63-84.

[2] Madhvanath, Sriganesh, and VenuGovindaraju. "The role of holistic paradigms in handwritten word recognition." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.2 (2001):149-164.

[3] Park, Jaehwa, VenuGovindaraju, and Sargur N. Srihari. "OCR in a hierarchical feature space." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.4 (2000): 400-407.

[4] Srihari, Sargur N., et al. "Individuality of handwriting." *Journal of Forensic Sciences* 47.4 (2002): 856- 872.

[5] Liu, Xia, and Zhixin Shi. "A format-driven handwritten word recognition system."*2013 12th International Conference on Document Analysis and Recognition*. Vol. 2. IEEE Computer Society,2003.

[6] Tappert, Charles C., Ching Y. Suen, and Toru Wakahara. "The state of the art in online handwriting recognition." *IEEE Transactions on pattern analysis and machine intelligence* 12.8 (1990):787-808.

[7] Bunke, H., SamyBengio, and A. Vinciarelli. "Offline recognition of unconstrained handwritten texts using HMMs and statistical language models."*Pattern Analysis*