International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 38003-38008
© Research India Publications. http://www.ripublication.com

A Unified System For Offline Handwritten Character Recognition And Language Interpretation

Magesh Kasthuri
Research Scholar, SCSVMV University, Kanchipuram magesh.kasthuri@wipro.com

Dr.V.Shanthi
Professor, Department of MCA, St. Joseph's college of Engineering, Chennai – 119 drvshanthi@yahoo.co.in

Abstract

Handwriting recognition principally entails optical character recognition. However, a complete handwriting recognition system also handles formatting, performs correct segmentation into characters and finds Off-line recognition.

Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. The idea of this article is to propose offline character recognition based on self-training and adjacent cell recognition.

Language identification and interpretation of handwritten characters is one of the challenges faced in various industries. For example, it is always a big challenge in data interpretation from cheques in banks, language identification and translated messages from ancient script in the form of manuscripts, palm scripts and stone carvings to name a few.

Therefore, there is need for greater accuracy in offline handwriting recognition of such handwritten text. Hence displaying the confidence of recognition helps the user to decide if this can be taken as acceptable threshold or improvise with further noise reduction or manual correction process.

Keywords: Character recognition, noise reduction, pre-processing techniques in character recognition, pattern matching, strokes, fixed-language, training neural networks, Feature Extraction, SVM Classifier

Introduction

Off-line character recognition often involves scanning a form or document written sometime in the past. This means the individual characters contained in the scanned image will need to be extracted. Tools exist that are capable of performing this step however, several common imperfections in this step. The most common being characters that are connected together are returned as a single sub-image containing both characters. This causes a major problem in the recognition stage. Yet many algorithms are available that reduce the risk of connected characters.

Neural network recognizers learn from an initial image training set. The trained network then makes the character identifications. Each neural network uniquely learns the properties that differentiate training images. It then looks for similar properties in the target image to be identified. Neural networks are quick to setup; however, they can be inaccurate if they learn properties that are not important in the target data.

A stroke is not limited to a continuous line segment. A stroke may also include a portion of a character that has a discontinuity in its representation. For example, an English alphabet 'i' may also be considered as a single stroke according to some embodiments in spite of a discontinuity in its representation because there is no sudden change in angle in any portion of this alphabet.

Handwritten character recognition using Soft computing methods like neural networks is always a big area of research for long time and there are multiple theories and algorithms developed in the area of neural networks for handwritten character recognition.

Problem description

Neural networks are recognized as one of the most effective artificial intelligence technology for pattern recognition. Superior results in pattern recognition can be directly applied to business applications in forecasting, classification and data analysis. Following are some benefits of this approach:

Automation: The performance of neural networks is highly automated and thus minimizes human interaction.

High Accuracy: Neural networks are able to approximate complex non-linear mappings.

Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing and noisy data.

In sync with reality: Neural networks are easily updated with fresh data, making them useful for dynamic environments.

Ability to filter knowledge from data: Knowledge from data can only be derived through expert analysis.

Neural applications use complex mathematical algorithms to process vast amounts of data and categorize them as a human does. But neural applications can examine far more data in less time than a human can.

Language identification and interpretation of handwritten characters is one of the challenges faced in various industries. For example, it is always a big challenge in data interpretation from cheques in banks, language identification and translated messages from ancient script in the form of manuscripts, palm scripts and stone carvings to name a few.

Neural network recognizers learn from an initial image training set. The trained network then makes the character identifications. Each neural network uniquely learns the properties that differentiate training images. It then looks for similar properties in the target image to be identified. Neural networks are quick to setup; however, they can be inaccurate if they learn properties that are not important in the target data.

A stroke is not limited to a continuous line segment. A stroke may also include a portion of a character that has a discontinuity in its representation. For example, an English alphabet 'i' may also be considered as a single stroke according to some embodiments in spite of a discontinuity in its representation because there is no sudden change in angle in any portion of this alphabet. Therefore, there is need for greater accuracy in offline handwriting recognition of such handwritten text. Hence displaying the confidence of recognition helps the user to decide if this can be taken as acceptable threshold or improvise with further noise reduction or manual correction process.

Feature extraction works in a similar fashion to neural network recognizers however, programmers must manually determine the properties they feel are important.

Some example properties might be:

- Aspect Ratio
- Percent of pixels above horizontal half point

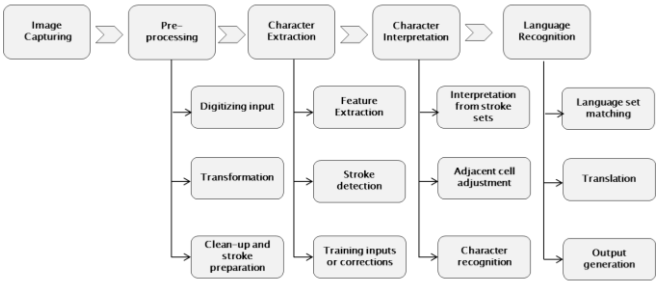


Fig.1 Proposed system for Handwritten Character Recognition

How this system differs from existing algorithms

Regarding the real-time application of offline handwritten character recognition techniques, here is the example where Handwriting Recognition techniques can prove to be solution for enhancing banking which satisfy the Visually Challenged person and may prove to be best banking solution. In order to build loyalty and drive profitability, banks need to offer a non-stop interactive banking environment.

To achieve this, banks need to increase their business agility by anticipating customer needs and offer an engaging user experience. The automatic processing of bank cheques involves extraction and recognition of handwritten or user entered information from different data fields on the cheque such as courtesy amount, legal amount, date, payee and signature. Hence, Automatic bank cheque processing systems are needed not only to counter the growing cheque fraud menace but also to improve productivity and allow for advanced customer-services.

The Automatic Cheque Processing System employed in Next Generation Banking may prove to be novel technology gaining the customers satisfaction (good - will) and recognition of skilled forgery by efficient validation techniques.

Quantitative and Qualitative Benefits

The suggested handwritten character recognition system applied as a banking technique employed in the modern banking system proves to be efficient time saving system. Also, the automated validation system using such artificial intelligence mechanism minimizes the forgeries and proves to be cost effective system to the bank. In such cases, Manual Efforts spent in validation of cheques can be reduced and also improves the performance of the bank by dedicating the valuable time for other business activities involved in Banking.

Further to this, ATMs enhanced with Handwriting Recognition technique for Customer Validation can prove to be effective user friendly technique for visually challenged person.

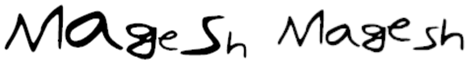
Training methods

Automatic recognition of handwritten dates present on bank cheques is also very important in application environments where cheques will not be processed prior to the dates shown. In countries like India, a cheque cannot be processed after six months of the date written on it. Verification of the hand printed signature present on a paper cheque is the most important challenge as the signature carries the authenticity of the Cheque. Character recognition, the process of converting the gray or binary images that contain textual information to electronic representation of characters that facilitate post-processing including data validation and syntax analysis is done to preserve the authenticity and avoid forgeries.

The first step is to obtain the image of the paper cheque using a scanner. Image acquisition involves acquiring the image of a form in color, gray level, or binary format. Preprocessing and segmentation modules follow the image acquisition step.

The verification and recognition of different information present in the cheque are done after the extraction phase. Nowadays while processing a cheque, banks are interested to read automatically as much information as possible from the document. This may include the payee-name, payer's address, and payer's account number, name of the issuing-bank and code lines. Main aspects related to preprocessing involves quality assurance, authentication, Binarization, skew correction as shown in below figure, slant correction and normalization are presented.

This process is also called Noise reduction or make-over of the character to be recognized. After pre-processing, it is necessary to perform the extraction operation of different handwritten fields prior to their recognition.



Skewed Signature Skew Corrected Signature

Fig.2 Pre-processing the text before character recognition

Segmentation of characters

The extraction of data present in cheques involves feature extraction of the information available in paper cherubs. If the goal of feature extraction is to map input patterns onto points in a feature space, the purpose of classification is to assign each point in the space with a class label or membership scores to the defined classes. Hence, once a pattern is mapped (represented), the problem becomes one of the classical classification to which a variety of classification methods can be applied to recognize the characters.

In American bank cheques, the extraction of legal amount includes the extraction of 'dollar' and 'cent' portions of the amount. This process is initiated by searching for a long horizontal line, which is usually written after the dollar part. If such a line is not present, the right side of the image is searched to find dashes and slashes to locate the cent portion. In, the courtesy amount recognition starts with the localization of handwritten numerical string based on the location of the courtesy amount.

In India, the courtesy amount is located on the right half of the cheque. A box is detected by identifying the cross-section points where horizontal and vertical lines meet. Another method proposed for extraction in uses fuzzy membership values, entropy, energy and aspect ratio as features, which are fed into a fuzzy neural network (FNN) for the identification of a field.

In English texts, words are separated by apparent space, but the letters within a word are not well separated, so words can be considered as natural units to recognize. In Chinese and Japanese texts, a sentence separated by punctuation marks is an integral unit because there is no difference between the inter-word and inter-character gaps. Word or sentence recognition, or generally, character string recognition, faces the difficulty of character segmentation: the constituent characters cannot be reliably segmented before they are recognized.

Experimental setup

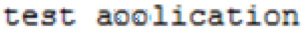
From a hard copy document, image will be extracted for offline character recognition. There are quite a few conventions in determining the input mode of such offline character recognition viz:

- complete document read / scan
- sequential reading (word / sentence wise) from the document (scan)
- Reading character / word / sentence directly from the digital image of the document

Consider a scanned text "test application" which is fed to the system for recognition as follows:

test application

The system first learns the strokes by itself and maps to the character set (alphabet series) it stores in the knowledge base. Hence before training (including pre-processing and noise reduction) it is able to recognize some of the characters like


test aoolication

The crucial step of creating index structure for all stroke sets from various character systems to be done like below

Table 1 Data model of storing trained data set

Language	Character	Handwriting style 1			Handwriting style 2			Handwriting style 3		
English	A	EA1	EA2	EA3	EA4	EA5		EA6	EA7	EA8
	B	EB1	EB2	EB3	EB4	EB5	EB6	EB7	EB8	
	C	EC1	EC2	EC3	EC4	EC5	C6	EC7	EC8	EC9
	D	ED1	ED2		ED3	ED4		ED5	ED6	EC9
Tamil	அ	TA1	TA2	TA3	TA4	TA5	TA6	TA7	TA8	TA9
	ஆ	TAA1	TAA2		TAA3	TAA4	TAA5	TAA6	TAA7	TAA8
	இ	TE1	TE2	TE3	TE4	TE5	TE6	TE7	TE8	TE9
	ஈ	TEE1	TEE2		TEE3	TEE4		TEE5	TEE6	TEE9
Hindi	अ	HA1	HA2	HA3	HA4	HA5		HA6	HA7	HA8
	आ	HB1	HB2	HB3	HB4	HB5	HB6	HB7	HB8	HB8

The final step is to store the pattern sets in the system for reference and recognition processing using the index.

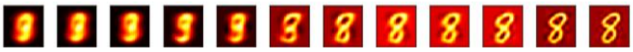


Fig.4 Example showing adjacent cell reading from a matrix of strokes

The above mapping table illustrates characters 'A', 'B', 'C', and 'D' of English and each character is mapped to multiple sets of second parameters such as set A1, set A2, set A3 etc. Each of these sets includes one or more parameters associated with a stroke into which that character can be segmented. Thus, each of these sets may individually represent a different stroke associated with the character. For example, handwritten character 'A' may be represented by three strokes respectively associated with set A1, set A2, and set A3 according to a handwriting style 1 of a first user.

Similarly, handwritten character 'A' may be represented by two strokes respectively associated with set A4 and set A5 according to handwritten style 2 of a second user. Here, handwriting style 2 may represent A in a different manner than handwriting style 1 because of different handwriting styles of two users. Therefore, the processor, while storing these second parameters, may interpret the character to be including three strokes according to style 1 and two strokes according to style 2.

Performance study with other methods

The system proposed by Mukarambi, Gururaj, et al[6] is an offline character recognition system with English and Kannada trained datasets and uses Feature Extraction and SVM Classifier method for recognition. When compared to the experimental setup of the proposed system of this article, Accuracy gets reduced in this system with different characters of similar shapes as experimented with sample size of 2550 with accuracy recorded as 83.02%.

There is another system proposed by Al-Marakeby, A., et al[7] which is based on AOCR, QDF based Zernike moments for offline Arabic character recognition and tested with a training sample of 23 providing recognition rate of 96.19% which is much better and consistent than first results but this system uses specific hardware designed AOCR which makes the system not a generic purpose solution.

On the other hand, Liu, Cheng-Lin, et al[9] proposed a Chinese handwritten character recognition which support offline recognition based on Normalization and Feature Extraction and online recognition based on MQDF, NPC, DFE, DLQDF and the recognition rate are 92.18% and 95.77% respective for offline and online recognition for a test experiment containing 224k samples. But the offline system proposed here is not appreciated in real-time due to higher cost and time for training and online recognition is observed to have mis-written characters, cursive shapes reduces accuracy.

There is another system proposed by Kumar, Munish, et al[1] which contains multiple recognition methods for Gurmukhi characters like SVM and Parabola curve fitting based features and SVM and power curve fitting based features in which the former produces result with 94.29% recognition rate with the later gives 97.14% for a sample of 3500 characters. In this system, recognition can be increased with 5-fold validation technique. Also, [1] alters the technique with k-NN classifiers in place of SVM as k-NN classifiers and Parabola curve fitting based features and k-NN classifiers and power curve fitting based features which has recognition rate of 84.17% and 90.06% respectively for same samples. The drawback on this k-NN based system is that Recognition accuracy is low in parabola curve fitting and Higher memory for training data and classification speed in power curve fitting.

Favata, John T, et all[12] proposes a offline system for English characters using GSC Classifier which gives 86.62% results on a 7000 character samples and the observation is that it has Low recognition rate as compared to higher training cost.

Finally, in this experimental analysis and comparison to observe the benefits of proposed system, there was another system taken for observation which was proposed by Bai, Jinfeng, et al.[2] which is a offline Chinese recognition system based on Feature Extraction, Stroke based recognition and produces accuracy of 85.44% for a sample set of 12700 characters. This is a Low recognition rate and higher training cost.

Now, the proposed system as compared to these is an multi-language recognition system as the recognition is done based on character sets for each group of strokes (segments) as explained in Table 1. Also, it is self training based system and works on adjacent stroke based recognition and hence recognition rate would get improved by practice and no additional cost or time involved in training and corrections.

The system has a pre-processing stage which helps in normalizing the segments and makeup of characters in order to achieve higher recognition rate. Metrics on individual character recognition Accuracy Ratio is shown below:

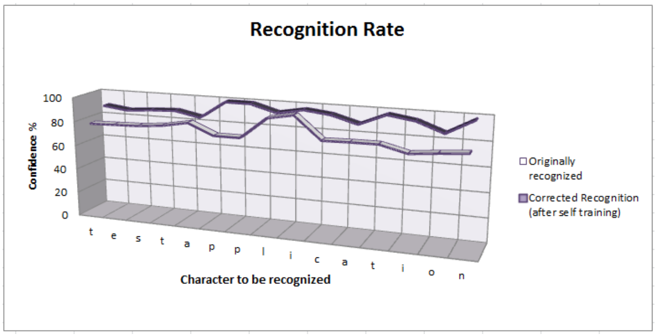


Fig.5 Metrics – Accuracy Ratio

With this setup, when a distributed variance of input text are fed to the system for recognition, there would be interesting statistics based on number of lines to scan and time taken in producing the result co-related with accuracy level of the recognition.

For example, the knowledge base may be populated with 15 million or more unique patterns of a lower case machine print letter `a` and may have 25 million or more unique patterns of an upper case machine print letter `R`. The recognition engine used in these systems on a high level uses this massive knowledge base of stored patterns to compare or match against the letters and words that reside on an image file or document image being recognized. Huntington et al.[5] disclose a knowledge base of pre-stored pixel/dot patterns for a vast array of digitized images, scanned images, characters, and words. The pattern discloses different variations of the complete character. However, they fail to disclose or suggest segmenting the image representing the character into the one or more first strokes which is the key difference in the proposed system.

As *Huntington* fails to disclose or suggest segmenting the image representing the character into the one or more first strokes, therefore, Huntington also fails to disclose or suggest a set of first parameters associated with each of **segmented** the one or more first strokes which is the key for better performance and higher recognition rate of the proposed system.

Further, Zhang[7] and Chen[8] fail to cure the above noted defects of Huntington. Chen idea is collecting handwritten characters and performing handwriting recognition based on parameters calculated from strokes of the handwritten characters. Stroke start and end events are identified and stroke parameters are calculated from coordinates of the stroke start and end events. One or more candidate characters are identified based on the stroke parameters.

As noted above, Chen merely discloses identifying the start and end events of a stroke. However, Chen fails to disclose or suggest segmenting the image representing the character into the one or more first strokes. As *Chen* fails to disclose or suggest segmenting the image representing the character into the one or more first strokes, they fail to disclose or suggest a set of first parameters associated with each of **segmented** the one or more first strokes which is another key process of the proposed system.

It should be apparent to a person skilled in the art that this system is not limited to storing second parameters associated with the strokes of only English characters. The scope of this system also includes storing second parameters associated with strokes of characters of various known languages, number systems, special characters, or symbols. It should be further be apparent that the number of stored second sets of parameters per handwriting style and the number of stored handwriting styles is not limited as illustrated here and can be more or less than this number.

Conclusion

Handwriting recognition principally entails optical character recognition. However, a complete handwriting recognition system also handles formatting, performs correct segmentation into characters and finds Off-line recognition. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand "printed" text. There is no OCR/ICR engine that supports handwriting recognition as of today.

Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles.

The approach towards creation of AI in machines has spawned numerous techniques. Some of these techniques are discussed in this paper. Statistical Learning Algorithms are based on slow learning algorithms that draw inferences from previous data. They however do not have the capability to make any differentiation between separate behavior patterns. The results however can be applied to a number of applications concerning quantitative data.

They can be used to:

- ✓ Cluster descriptors in terms of a relatively small set of characteristics
- ✓ Test hypotheses concerning differences among populations
- ✓ Perform trend analysis, as in the case of time series analysis and construct correlations among sets of variables

There are two major techniques and devices in this regard called OCR and ICR. OCR which is termed as Optical Character recognition is based on legacy model where character recognition is based on optical scanning. ICR which is termed as Intelligent Character recognition is a modern engine which has different pattern of character recognition including but not limited to handwritten character recognition.

There is no OCR/ICR engine that supports handwriting recognition as of today. This paper work proposes a unified system capable of character recognition with mixed language content without time consuming preparation steps of training the networks. Rather, this system proposes a new concept called self-training where the system trains itself using Artificial intelligence by applying Genetic algorithms to build neural network system on its own by repeatedly processing the character sets.

A further study in this research topic is preparing unified system for language identification based on Language detection libraries from the interpreted characters. This can be further integrated to Speech recognition system (for example: Text to speech recognition engine) to read out the interpreted text which can be a excellent use case for Visually challenged people where they struggle (or manual need to support) to get proper interpretation of text books and pdf documents.

Acknowledgement

This concept is a sub-section of US/India patent filed concept and acknowledged by the patent (Indian Patent Journal Issue 47/2013 dated 22/11/2013). All content and idea are copyrighted. The patent is published by USPTO (pre-publishing) as U.S. Patent Application 14/146,213 Publication number US20150131912 A1.

References

[1] Kumar, Munish, R. K. Sharma, and M. K. Jindal. "Efficient Feature Extraction Techniques for Offline Handwritten Gurmukhi Character Recognition." National Academy Science Letters 37.4 (2014): 381-391.

[2] Bai, Jinfeng, et al. "Chinese Image Character Recognition Using DNN and Machine Simulated Training Samples." Artificial Neural Networks and Machine Learning–ICANN 2014. Springer International Publishing, 2014. 209-216.

[3] Mitra, Chandana, and Arun K. Pujari. "Directional Decomposition for Odia Character Recognition." Mining Intelligence and Knowledge Exploration. Springer International Publishing, 2013. 270-278

[4] Mori, Minoru, Seiichi Uchida, and Hitoshi Sakano. "Dynamic Programming Matching with Global Features for Online Character Recognition." ICFHR. 2012.

[5] Huntington, Stephen G., Bevan Rowley, and E. Derek Rowley. "Method of massive parallel pattern matching against a progressively-exhaustive knowledge base of patterns." U.S. Patent No. 8,391,609. 5 Mar. 2013.

[6] Abdulkader, Ahmad A., et al. "Template-based cursive handwriting recognition." U.S. Patent No. 7,369,702. 6 May 2008.

[7] Zhang, Qi, et al. "Stroke segmentation for template-based cursive handwriting recognition." U.S. Patent No. 7,302,099. 27 Nov. 2007.

[8] Chen, Yen-Fu, and John Dunsmoir. "Method and apparatus for performing handwriting recognition by analysis of stroke start and end points." U.S. Patent Application 10/756,918.

[9] Liu, Cheng-Lin, et al. "Online and offline handwritten Chinese character recognition: benchmarking on new databases." Pattern Recognition 46.1 (2013): 155-162.

[10] Yang, Yang, Xu Lijia, and Cheng Chen. "English character recognition based on feature combination." Procedia Engineering 24 (2011): 159-164.

[11] Aghav, Sushila, and S. S. Paygude. "Computer Assisted Printed Character Recognition in Document Based Images." Procedia Engineering 38 (2012): 3222-3227.

[12] Favata, John T., and Geetha Srikantan. "A multiple feature/resolution approach to handprinted digit and character recognition." International journal of imaging systems and technology 7.4 (1996): 304-311.

[13] Kasthuri, Magesh, V. Shanthi, and Venkatasubramanian Sivaprasatham. "Mixed Language Based Offline Handwritten Character Recognition Using First Stroke Based Training Sets." International Journal of Image Processing (IJIP) 8.5 (2014): 313.

[14] Kasthuri, Magesh, and V. Shanthi. "Techniques in Neural Network Recognition and its Relation to Brain Theory." Artificial Intelligent Systems and Machine Learning 4.7 (2012): 459-465.

[15] Kasthuri, Magesh. "Systems and methods for offline character recognition." U.S. Patent Application 14/146,213 Publication number US20150131912 A1.