A Survey on Feature Extraction Techniques for Handwritten Text in Various Scripts

Pooja Sharma
Assistant Professor, D. A. V. College, Abohar,
Punjab, India
poojasharma8aug@gmail.com

Abstract

Optical Character Recognition (OCR) System is an interesting area to recognize pattern of the documented image to make it in editable form, where the given text can be in hand written or printed form. There are so many recognition systems developed for various scripts used in the world and taking input in online and offline modes, with considerable efficiencies. These systems are applicable in the various fields like IT systems, Banking, Postal Sector and Education for automated information retrieval and digitization. In this paper, I present a survey on the techniques for recognizing handwritten as well as hand printed documents and also focus on the Feature Extraction phase with the corresponding classification technique.

Keywords: OCR, Classification, Feature/ Extraction.

I. INTRODUCTION

The domain of computer vision and Digital Image Processing give a way for devising an approach for development of an OCR system meant to understand the digital image and infer from it the editable document required for various applications such as automated data entry and retrieval, bank cheque analysis, hand written pictogram interpretation, hand written formulae retrieval, address retrieval and verification and many more. Various researchers all over the world have proposed and applied methodologies to serve solutions to these problems. For many languages and modes of input for all these applications, the OCR system holds to be robust. However more efficient systems may also be developed, keeping in view the requirements and nature of input documents for certain specific applications such as for OMR forms processing, where hand printed data can be recognized by machines in addition to the recognition of encircled bubbles.

A. Background

Background of the research work carried out in field of Hand written document image recognition goes back to decades. The history of OCR in field of Handwritten Text Recognition has been discussed by J.R Prasad and Kulkarni [1] as trends in Handwriting recognition. The earliest work in this domain according to the mentioned survey work [1], has been reported to be done in the early sixties and seventies .Character images were then modelled as being composed of blocks defined geometrically by the co-ordinates of their vertices and the specification of edge information. Later integrated segmentation and interpretation systems came into being.

Then there came into existence, the algorithmic and computational techniques to process the image at high level (edge finding, region growing and segmentation).Further improvements were made to these techniques to obtain more efficient algorithms for classification such as Hidden Markov Models (HMM), Dynamic Programming, Neural networks etc.[1]. Thus the trend moved on towards devising automated techniques for

recognition, incorporating efficient algorithms for pre processing, feature extraction and classification. Also use of combination of recognizers, use of lexicons, dictionaries and language models [1] is being introduced to achieve higher recognition rates.

B. Handwritten v/s Hand printed Text

Certain terms such as handwritten, hand printed, on-line and off-line, often misinterpreted, need to be clarified before moving further. The term Hand printed document image depicts that input image comprises of mono-spaced characters, i.e. each character has the same distance with all the other characters. In handwritten document, we have characters in the cursive form whereas Hand printed input document has some uniform height or width. Thus, as many variable elements are not introduced as in straight handwriting (unconstrained). The other terms being off-line and on-line recognition [2] modes of obtaining input for recognition can be explained as follows. On-line document image recognition implies to storing in order the 2-dimensional co-ordinates for the successive points of writing as a function of time, thus the spatio-temporal information such as order of strokes done by the writer, information about pressure and angle of the pen is readily available. In case of off-line document where acquisition is done prior to recognition, the image consists of the complete data to be written. Here the spatio-luminance of the given image is analyzed. Therefore, more challenges would be present while recognizing documents in offline mode since we have only static information about the document.

II. FEATURE EXTRACTION-A CRUCIAL PHASE IN THE OCR PROCESS

Firstly, pre-processing phase is completed, then desired level of segmentation (regarding line, word, character or symbol) is to be achieved, we use some feature extraction techniques on the segments for obtaining features, then classification and after that post processing techniques are to be implemented. It is necessary to focus on feature extraction because it has a high impact on the recognition system's efficiency. Anil K. Jain, Taxt [3] emphasizes on the fact that feature selection for a feature extraction method to acquiring high recognition performance [3]. Feature extraction means extracting from the raw information for classification purposes. Thus, an appropriate feature extraction technique selection needs to be applied according to the given input very carefully. Also, efficiency of classification and therefore, that of recognition depends upon the feature extraction. As we obtain an n-dimensional feature vector from this phase that is fed into the classifier for further processing, feature set hence determined needs to be optimal enough to give efficient results, i.e. it must take into consideration the curse of dimensionality [3] as mentioned in the survey work by A.K Jain and Taxt.. Taking the consideration of all these factors, it becomes necessary to look at the different techniques for feature extraction in the given domain for covering vast possibilities of cases. Therefore we move forward to have a look over the techniques developed till date for handwritten and hand printed character image recognition in off-line mode, where we cover various languages and scripts-Assamese, Persian, Thai, Devanagari, Chinese and Roman characters and digits.

III. FEATURE EXTRACTION TECHNIQUES

The basic classification can be done mainly on the basis of representation of input document as on-line and off-line, where in each category, we may have numerous techniques-statistical, structural, rule based etc. for both hand printed and handwritten document recognition in various language scripts. The various techniques are as follows-

A. Hand printed Numeral Recognition

Various scripts discussed for numeral recognition are- Roman, Assamese and Persian. One of the techniques proposed by J.R Parker [4] to recognize hand printed digits is to use scalable vector templates which generate templates with same scale and line width attributes as an arbitrary image[4].There was the problem to match „1" or any other object that had extreme height to width ratio. Shridhar et al [5], had used a combination of global- left and right profiles of external contours, difference of left and right profiles and local features- width of character, ratio, location of maximum ad minimum and right, left peaks of the digit. A syntactic recognition algorithm was implemented by formulation of production rules [5], resulting into overall accuracy as 99%. Heutte et al [6] used seven features- both statistical and structural, combined to represent the numerals. These were intersection with straight lines, holes and concave arcs, invariant moments, end points and junctions, extrema, profiles, projections [6]. Statistical classifier [6] was employed and 90.8% accuracy was achieved. Liu et al [7] made use of variants of directional features- chain code, gradient feature with Sobel and Kirsh operators, and peripheral direction contributivity (PDC)[8] in complementation with profiles structural features[9] and concavities structural features[10] to compose ten feature sets[7]. Classifier system implemented by the authors [7], composed of eight classifiers (RBF, PC, k-NN, LVQ, DLQDF, MLP, SVC-poly, and SVC-RBF)[7] to give the highest accuracy as 99.58% when using SVC with RBF kernel (SVC-RBF)[7]. Koerich [11] have also made use of profiles as complementary features for recognition of hand-printed pattern. G. Siva Reddy et al.[12] achieved remarkable recognition rates 97.6% for offline Assamese handwritten numerals using VPP-HPP, Zonal DCT, CCH, Pixel level features and a combination of all the four features[12] to represent features and vector quantization[12] for modelling[12]. Omid Rashnoodi et al.[13] have performed offline recognition of handwritten Persian digits using five feature sets, composed of-variance, co-variance, central moments, median and number of pixels per each square[13] and two classifiers-SVM[13] and k-Nearest Neighbour[13] to give 91.3% and 92% as recognition rates, respectively.

B. Chinese and Thai Characters

Another research methodology for feature extraction in Chinese hand printed characters was to extract features using statistical methods. Dominant Point Method [14] was the most popular for the said problem, for which several algorithms had already been developed. However, the authors [14] felt a need to detect only the points with a very sharp curvature. The authors had followed the Rosenfeld-Johnson algorithm[14],whose basic concept had been stated as "to calculate curvature of each point in line, then the points with local maxima in curvature are designated as dominant points"[14].The authors had formulated six equations to denote the six primitives. The recognition rate [14] achieved was 84.45%.This methodology worked on the need to break traced lines into segments for easier recognition. This algorithm was affected due to irregularities of lines; however authors had proposed and applied re-merging of line segments which resulted into success. Another technique for feature extraction was applied on Hand printed Thai characters [15], which combined both global and local features of the characters. Statistical Features were represented by global features to define the shape of the characters, which was further used to group similar shape characters together. Local features were used to represent the symbolic structures for the characters. The structural features identified were loops, end points, junction points and curls. Authors had improved efficiency for recognition of unconstrained (handwritten characters).Also this methodology resulted out to be robust and an improvement over recognition of structural feature extraction, with efficiency of 87.32% [15] for global and local features.

C. Devanagari Script

There has been considerable development in field of Indian language OCR, where recognition of Hand-printed characters in Devanagari script has been done by implementing a combination of features to yield efficient results. Features implemented and compared [16] in this paper are as follows-

- Kirsch Directional Edges
- Distance Transform
- Chain Codes (Freeman Chain Codes)
- Directional Distance Distribution
- Neighborhood Pixels Weight
- Total Distances in four Directions

The classifiers used were MLP and SVM [16], that was fed with multiple features and results were compared. It was noted that lowest efficiency was obtained by using Kirsch Directional edges and highest using Gradient. Overall recognition rate obtained was 94.3% using SVM (Support Vector Machines) [16]. Satish Kumar [17] has also thrown some light on the analysis of following mentioned features, in addition to the five features mentioned according to [16] .These features [17] are-

- Zoning
- Profiles
- Chamfer
- Histogram
- Crossings
- Chessboard
- Chain code Histogram
- Gradient (Sobel)
- Total Distances in 4 directions combined with Gradient-TdistGrd-200
- Neighbouring pixels Weight combined with Gradient-NpwGrd-200.

Evaluation has been done using three classifiers- SVM(RBF),k-NN and MLP[17].Best result 93.5% has been stated as obtained using Gradient (Sobel) -200 [17]with SVM as classifier. Various feature extraction techniques for hand printed Devanagari character recognition have been enlisted by R. Jayadevan et al [18], where efficiencies of each of the techniques have been compared. Likewise a wide variety of features that have been used for Devanagari numerals and characters both, have been covered by the authors [18], some of them are as follows-

- Structural
- Statistical
- Density features
- Moment features of right, left, upper, and Lower profile curves
- Descriptive component features
- Shape descriptors.

A comparative study has been done by the authors and feature extraction techniques with corresponding recognition results [18] of each of the techniques have been tabulated. Ashutosh Aggarwal, Rajneesh et al [19] had worked on recognition of offline Devanagari characters, obtaining gradient feature vectors and applying SVM (RBF) as the classifier to yield recognition rates as 94%.

D. English language Hand printed and Handwritten Characters

Gilewski et al [20] performed in-depth laboratory work to study techniques for Handwriting Recognition. The authors had applied the fuzzy feature extraction algorithm [20], with the main aim being to extract the defined segments from the character image. Multilayer feed forward Neural Network was used in the laboratory in order to realize the proposed algorithm. Rafael M.O. Cruz et al [21] applied Multiple Feature Extraction Techniques for Cursive Character Recognition while making use of an ensemble classifier for recognition. There were Nine feature sets which had been constructed. Other techniques [21] can be enlisted as follows-

- Structural Characteristics, where radial, horizontal and vertical histograms were computed.
- Modified Edge Maps
- Image projections
- Multi Zoning
- Concavities Measurement
- MAT-Based Gradient directional features
- Gradient Directional Features
- Median Gradient Features
- Camastra 34D Features

These nine feature sets [21] were given to the ensemble system of classifiers. The motivation authors [21] felt for building ensemble system were as follows- Errors made by classifiers with different feature extraction methods do not overlap and Divide and Conquer paradigm, i.e. using each feature set separately and combining their results would prove to be more efficient. Authors have presented a comparison of results obtained for each feature set and each classifier. However, the new technique of Modified Edge Map produced the best overall results, 84.37% [21]. Offline Cursive Character recognition was performed by Anshul Gupta et al [22] incorporating heuristics based segmentation algorithm. The recognition was performed at the word level which required segmentation of word into independent characters (uppercase and lowercase). Fourier Descriptors [22] were used in order to extract features, namely Fourier Magnitude [22] and Fourier Angle [22]. There were three networks considered as classifiers- Radial Basis Function (RBF), Multi-layer perceptron (MLP) and Support Vector machine (SVM).Ten variations of 26 word images were considered. The proposed system [22] achieved recognition accuracy of 98.74% on the training dataset using SVM as the classifier .As stated by the authors; it had outperformed RBF and MLP [22].

E. Assamese language

In [23], the recognition of online handwriting for Assamese language is done using both HMM and SVM modeling and their performance is compared. 181 different Assamese strokes are trained to generate Recognition models by using HTK (HMM Toolkit) and LIBSVM (SVM Toolkit). The developed stroke classifier gives average recognition accuracy rate of about 94 % in case of HMM and 96 % in case of SVM. The akshara level average performance rate is 84.67 % in case of HMM and 86.23 % in case of SVM. The SVM based system gives better performance than HMM based system by 2 % in stroke accuracy and then 1.56 % in akshara case. In [24], comparison of Assamese Character Recognizer is done using Stroke Level and Character Level Engines. Assamese is the major language of Assam. There are total 11 vowels, 41 consonants and 10 numerals in the Assamese language. An Assamese character set consists of one or more strokes. There are two approaches used in this paper for character recognition using HMMs. In the first approach, HMMs were trained on stroke level data and characters were recognized by comparing stroke sequences generated by HMM. The untrained character was recognized by adding their stroke combination in the reference set. In the second approach, HMMs were directly built on character level data. Variations in the stroke order and sequences to write a character, led to the difficulty in updating the reference set. In that case the character based method was used as it was more efficient. , the stroke based character recognition system gave an average accuracy of 43.76% and the average accuracy of the character recognizer using character HMMs was obtained as 82.96%.The summarization of the above mentioned feature extraction and classification techniques with recognition rates achieved in each case have been tabulated in Table I.

Table I. Feature Extraction techniques and corresponding classifiers

S. No.	Languages	Feature Extraction Technique	Classification Technique	Result
I.	Hand printed Numerals(Roman digits) [4]	Vector Template Representation	Template Matching	94.3%(Average result)
	Persian Digits [13]	Five Features" set	SVM	91.3%
			k-NN	92%
II.	Chinese Characters[14]	Freeman chain codes	Statistical Discriminant Analysis	84.45%
III.	Thai Characters[15]	Local-symbolic and Global-statistical	NLP Neural network Models	87.32%
IV.	Devanagari Characters [16], [17]	TDIST and four other techniques	MLP and SVM	94.3% (Overall using SVM)
		Gradient(Sobel)	SVM(RBF)	93.5%
V.	English language(Roman Script) [21], [22]	Nine features" set	Ensemble System	84.37%(Modified Edge Maps as Features)
		Fourier Descriptors	SVM	98.37%
VI.	Assamese character database [23], [24]	Stroke feature using HTK (HMM Toolkit) and LIBSVM (SVM Toolkit)	HMM and SVM	For strokes-HMM (94%), SVM (96%) For aksharas-HMM

IV. CONCLUSION

The intent of covering these techniques was to have an overview about the existing feature extraction techniques and perform a comparative analysis of various such techniques. Much of work has been completed taking into account statistical and structural features, both independently and in combination as well. However, more work is required to be done using rule based methods for feature extraction, which may lead to even better results in future. It can be seen that use of profiles has been prevalent especially for recognition of roman digits. Statistical techniques like multi zoning and edge maps yielded good efficiencies for the hand written Roman alphabet recognition, for which the reason could be a large feature set is obtainable in these cases. For classification, the best efficiencies have been determined by using intelligent systems, specifically SVM as the classifier. However the ensemble system also holds out to be a strong competitor for being chosen as the classifier. The techniques for feature extraction and classification hence discussed would prove to be helpful to develop a clear understanding of the concepts essential to build a recognition system for offline hand written documents, even applicable to scripts other than those discussed in this survey.

REFERENCES

[1] J. R. Prasad and U.V. Kulkarni, "Trends in Handwriting Recognition", IEEE, (2010).

[2] R. Plamondon and S. N. Srihari, "On-line and Off-line Handwriting Recognition-A Comprehensive Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No.1, (2000).

[3] I. D. Trier, A. K. Jain and T. Taxt, "Feature Extraction Methods for Character Recognition-a Survey", Pattern Recognition Vol.29, No.4, (1996), pp.641-662.

[4] J. R .Parker, "Vector Templates and Hand printed Digit Recognition", IEEE-(1994), pp.1051-4651.

[5] M. Shridhar and A. Badreldin, "A High Accuracy Syntactic Recognition Algorithm for handwritten Numerals", Proceedings of the IEEE International Conference on Systems , Man and Cybernetics, Vol. SMC-15, No. 1, (1985)

[6] L. Heutte, T. Paquet, J. Moreau, Y. Lecourtier and C. Olivier, "Combining Structural and Statistical Features for the Recognition of Handwritten Characters", ICPR, Vol. 19, (1998) pp. 629-641.

[7] C. L. Liu, K. Nakashima, H. Sako and H. Fujisawa, "Handwritten Digit Recognition: Benchmarking of State-of-the-Art", Pattern Recognition, No. 36, (2003), pp. 2271- 2285.

[8] A. de S. Britto Jr, E. Lethelier, R. Sabourin and F. Bortolozzi, "Improvement in handwritten numeral string recognition by slant normalization and contextual information". Proceedings of the Seventh International Workshop on Frontiers of Handwriting Recognition, Amsterdam, (2000), pp. 323–332.

[9] J.T. Favata, G. Srikantan and S.N. Srihari, "Hand printed character/digit recognition using a multiple feature/resolution philosophy". Proceedings of the Fourth International Workshop on Frontiers of Handwriting Recognition, Taipei, (1994), pp.57–66.

[10] J.T. Favata, G. Srikantan, S.N. Srihari, "Hand printed character/digit recognition using a multiple feature/resolution philosophy", Proceedings of the Fourth International Workshop on Frontiers of Handwriting Recognition, Taipei, (1994), pp.57–66.

[11] L. Koerich, "Large Vocabulary Off-line Handwritten Word Recognition", Ph. D. Thesis, Ecole de Technologie Superieure, Montreal - Canada, (2004).

[12] G.S. Reddy, P. Sharma, S.R.M Prassana, C.Mahanta , L.N. Sharma, "Combined online and Offline Assamese Numeral Recognizer" /978-1-4673-0816-8/12 IEEE (2012)

[13] O. Rashnoodi, A. Rashnoodi and A. Rashnoodi, "Offline Recognition of Handwritten Persian Digits using Statistical Concepts". International Journal of Computer Applications Vol. 53, No.8, (2012), pp. 0975 – 8887.

[14] A. Amin and S. Singh, "Machine Recognition of Hand Printed Chinese Characters" Intelligent Data Analysis (1997), pp. 101- 118

[15] I. Methasate, S. Marukatat, S. Saetang and T. Theeramunkong, "The Feature Combination Technique for Offline Thai Character Recognition System" IEEE, Eighth ICDAR, (2005).

[16] S. Kumar, "Performance Comparison of Features on Devanagari Hand Printed Dataset", International Journal of Recent Trends in Engineering, Vol.1, No. 2 , (2009).

[17] S. Kumar, "Study of Features for Handprinted Recognition", World Academy of Science, Engineering and Technology 60, (2011).

[18] R. Jayadevan, S. R. Kolhe, P. M. Patil and U. Pal, "Offline Recognition of Devanagari Script-A Survey", IEEE transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, Vol.41, No.6, (2011).

[19] A. Aggarwal, R. Rani, R. Dhir, "Handwritten Devanagari Character recognition Using Gradient Features", IJARCSSE Vol. 2, No. 5, (2012).

[20] J. Gilewski , P. Phillips, S. Yanushkevich and D. Popel, "Education Aspects-Handwriting Recognition using Neural Networks, fuzzy Logic" Pattern recognition and Information processing, vol. 1, (1997), pp. 39-47.

[21] R. M.O Cruz, G. D.C Cavalcanti, T. I. Ren, "Ensemble Classifier for Offline Cursive Character Recognition Using Multiple Feature Extraction Techniques". IEEE, (2010).

[22] A.Gupta, M. Srivastava and C.Mahanta, "Offline Handwritten Character Recognition" ICCAIE, (2011).

[23] D. Das, R. Devi, S.R.M. Prasanna, S. Ghosh, K. Naik. , "Performance Comparison of Online Handwriting Recognition System for Assamese Language Based on HMM and SVM Modelling", Computer Science and Information Technology, (2014).

[24] H. Choudhury, S. Mandal, S. Devnath, S. R. M. Prasanna and S. Sundaram, “Comparision of Assamese Character Recognizer using Stroke Level Engines”, Dept. of Electronics and Electrical Engg, IIT Guwahati, Guwahati, (2015).