Volume 6, Issue 9, September 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Systematic Review of Optical Character Recognition Techniques

Sandeep Kaur[*], Rekha Bhatia
CSE, Punjabi University,
Punjab, India

Abstract— Optical Character Recognition (OCR) is technique to convert the scanned handwritten, typed or printed document into computer readable form. OCR system helps to digitize old documents, recognizing vehicle number plate and have various surveillance and forensic applications. A systematic flow of OCR system can be divided into six phases: data collection, preprocessing, feature extraction, segmentation, recognition and post-processing. In this paper, a review of different techniques proposed for OCR system is provided phase by phase. This paper can help the researchers and designers to select appropriate technology as per required application.

Keywords— Optical Character Recognition, Handwritten Character Recognition, Pre-processing, Segmentation, Classification, Post-processing.

I. INTRODUCTION

This paper considers the procedure for the recognition of online handwritten characters by using the digitizer tablets or writing pads. Digitizer tablets are based on the resistive and analog to digital conversion techniques. These are used to measure the pen tip for the recognition of handwritten characters. The proposed procedure for the recognition of online handwritten characters contains the following phases: data collection, preprocessing, feature extraction, segmentation, recognition and post-processing. While doing the literature review, it has been noticed that we can perform the segmentation phase before or after the preprocessing phase. The input of one phase is the output of its previous phase. These phases are illustrated in Fig. 1.1. Subsections 2.1 to 2.6 discuss about these phases in detail. Subsection 3 shows the comparison of various handwritten character recognition techniques.

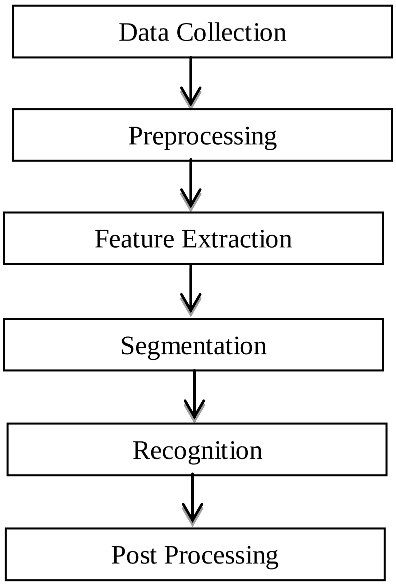

Fig 1.1 Phases of online handwritten character recognition.

II. LITERATURE REVIEW

A. *Data Collection*

Data collection is the first phase of the online handwritten character recognition system. In this phase, we collect the data for recognition of online handwritten characters by using the digitizer tablets or writing pads. This device contains a digitized pen. We can collect the data by collecting the coordinate points of this moving digitized pen sequentially. This pen basically performs two actions which are PenUp and PenDown. We can trace the connected coordinate points of the pen between the PenUp and Pendown and call it as stroke. Then these strokes are sampled at the constant rate. The common names and appearance of the writing pads are shown in Fig 2.1.

Crosspad | Tablet PC | Personal Digital Assistant

Fig. 2.1: Electronic devices used for capturing handwritten characters.

B. Pre Processing

In this phase, distortion and noise are removed from the collected data, which came with the data due to software or hardware limitations. These distortions and noise are presence of jitter in the text, irregular shape and size of the text, right or left bend in handwritting and missing coordinate points of digital pen during their collection. This phase mainly includes the five steps which are: centering and size normaiilization, smoothing, interpolating missing points, resampling of points and slant correction. These steps are shown in Fig 2.2.

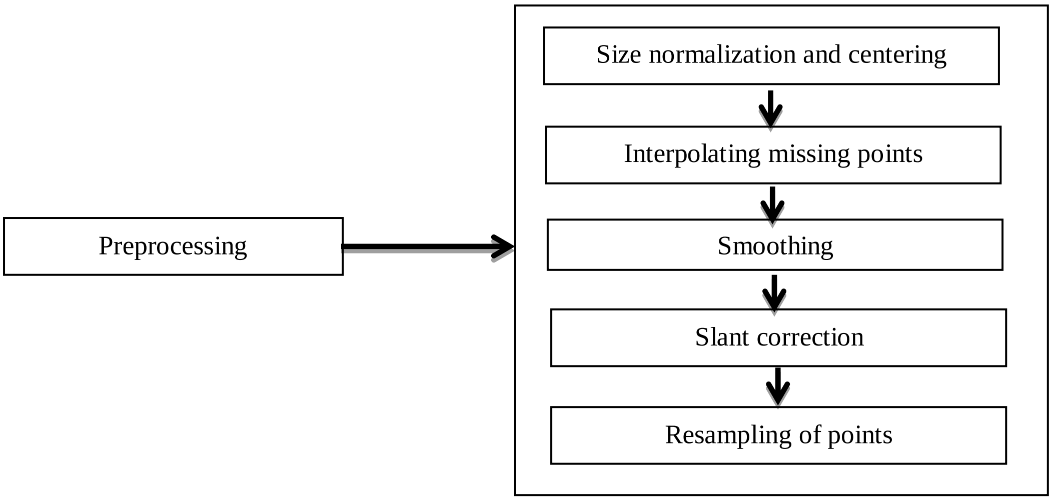
Preprocessing
Size normalization and centering
Interpolating missing points
Smoothing
Slant correction
Resampling of points

Fig 2.2: Five steps of Preprocessing phase

Beigi et al. ,1994[1] has discussed about the various size normalization techniques in his paper. Size normalization basically depends on the movement of the pen on the writing pad and centering is required during the movement of the digital pen on the writing pad along with the borders.

Unser et al., 1993[2] has discussed some methods, which helps in resolving high speed writing characters which may lead to the missing of coordinate points, which can be resolved by using various interpolation techniques.

Kavallieratou et al.,[3] 2002 has presented a procedure for the removal of the jitter from the input handwriting. To do this we performe the smoothing step, which basically averages a coordinate point by using its neighbour points.

Uchida et al., 2001[4] has presented a procedure which helps in resolving the shapes of characters for identifying them. Slant normalization is performed for the normalization of the shape of input character’s components because handwritten characters are mainly slant or italic.

Plamondon et al.,1993[5] has discussed about the resampling of coordinate points in his paper. This is the fifth step of pre-processing and it calculates the new points on basis of original collected points for the ease in recognition of handwritten characters.

C. Feature Extraction

This phase plays very important role in achieving the high character recognition rate by selecting the suitable and correct features of the handwritten characters. Selection of correct features also reduces the complexity of classification. Features of characters vary as we shift from one script to another. There are various feature extraction techniques developed for selecting the suitable or correct features by many authors.Mainly, features are categorized into two classes: high level and low level class. Features of high level class gives useful information such as dots, headline, straightline and these are derived from the features of low level class. Low level features give information about slant, position, area and direction. But there is no specific method for extracting features of particular script.

Rocha et al.,1994[6] has proposed a system for feature extraction which reduces the demnsions of the character representation. This system gives information about the shape of character, their interrelations and specific features of characters.

Hu et al. (1997,2000)[7][8] has proposed a system in which high level features are combined with low level features on simple points and these are able to cover a huge amount of input pattrens. Also these features have invariance property which is used for normalizing the curvature of features

Verma et al. ,2004[9] has presented a technique for feature extraction from the online handwritten character recognition, which works on the basis of directional, structrual and zoning characterstics information to generate a single feature vector. This method is able to extract features from raw data independently without changing the size of the characters.

Joshi et al.,2005[10] has proposed a system for the recognition of the online handwritten characters. The system uses the characters of Devanagiri script for the recognition. It recognize the characters on the basis of their structural features and achieve the 94.49% recognition rate.

Schlapbach et al.,2007[11] has presented a system independent from the text and language for identifying the online handwritten characters of different writers. Different types of feature sets are extracted from handwritten characters and for the distribution of the features Guassian mixture models are applied on the feature sets.

D. Segmentation

In this phase data is represented in the form of strokes or character, so that we can study the nature of characters or strokes individually.

Tappert et al., 1990[12] has divided the segmentation into two types: internal segmentation and external segmentation.Where external segmentation is performed before the recognition of the characters and also it saves a lot of computation and makes the job of recognizer much easier and simple.

Yanikoglu et al., 1998[13] has proposed a new algorithm for the segmentation of handwritten characters which works under the guidance of global characterstics of the characters of the handwritten text. They evaluate a cost function which provides the successive points of segmentation. The cost for the segmentation at a point is calculated by the sum of weighted values of four features at that point. Linear programming approach is used to calculate the weights of features.

Blumenstein et al.,1999[14] has given a new algorthm for segmentation for handwritten word recognition system. This algorithm works on artifical neural network. Intially it is trained by using the global features of over segmented handwritten word images and then the test images of handwriiten words are given as input to the algorithm.which were get recognized with the help of the trained data.

Plamondon et al.,2000[15] has provided a survey on both online and offline handwritten recognition system. In which it is observed that study of segmentation is more beneficial in offline handwriting recognition rather than online handwriting system. In both the online and offline handwriting recognition system strokes or characters are identified in world level segmentation.

E. Recognition

In this phase characters are recognized by using some methods which aremainly divided into four types: neural network, syntactical and structural,Statistical and elastic matching.These methods are discussed in subsections 2.5.1 to 2.5.4.

1) Neural network methods

These methods works on neural networks which consists of parallel computing systems having a large number of inter connected processors.

Matic et al. ,1993[16] has proposed a writer adaptable system for online character recognition. It works on the basis of time delay neural network, which is first trained by using the handwriting of many writers to recognize characters. The system is speed and memory efficient.

Cho ,1997[17] has proposed three different neural network classifier for the recognition of complex patterns. These classifier includes structure adaptive self organizing classifier,HMM hybrid classifier and multiple multilayer perceptron classifier. These classifier are used to solve the unconstrained problems on the handwritten numerals.

Kimura et al. ,1997[18] has proposed a two-stage hierarchical system for recognition of handprinted kanji characters. This system contains artifical neural network and a statistical pattern recognition module for the recognition of characters of a large number of categories including some similar type of category sets.

Yaeger et al.,1998[19] has proposed a handwritten character recognition system. The proposed system works by using the neural network techniques. For the recognition of characters multi layer perceptron is used by this system and it gives better results.

Jaeger et al. ,2001[20] has proposed a system for online handwritten character recognition which works on the basis of a multistate time delay neural network. It contains a hybrid architecture with the combination of features of both HMMs and artifical neural networks. The main features of multistate time delay neural networks are its nonlinear time alignment procedure and time-shift invariant architecture.

Shintani et al. ,2005[21] developed a small scale character recognition system. It works on four-layered neural network model, which is used to recognize the simple characters.For recognization it uses the patterns which are transformed by affine conversion.

2) Structural and Syntactical Method

Structural and syntactical methods are used to recognize the handwritten characters or patterns.These methods mainly consider the structure and grammar of the characters or patterns.

Connell et al., 2000[22] has described that structrual recognition gives information about the generation of the structure of character through some primitives, which is useful to identify the characters and syntactic pattern recognition is used to recognize the connection between different structural patterns as a syntax of a language. Thus, we can recognize a large number of patterns by using a some grammatical rules and primitives.

Nouboud et al., 1991[23] has proposed a a real-time constraint-free character recognition system by using a structural approach for the recognition of the online handwritten characters. In the proposed system, a chain code is used for the representation of the characters and classification is done on the basis of usage of a dedicated processor for the string comparison.

Bontempi et al.,1994[24] has proposed a procedure for character recognition in which genetic algorithm is used as a learning system for recognition of the characters and a string matcher is used for the classification of the characters.

Duneau et al. ,1994[25] has proposed a system which is dedicated for the recognition of the cursive characters of the English language written on the digital writing pad. It uses an analytical approach for the recognition which localize the characters of the words on the basis of the prototypes of characters.

3) Statistical Method

In Statisttical approach, features are selected for the classification of pattrens or characters and statiscal methods works on the basis of prior probability of the classes. These are classified into two categories: non-parametric and parametric method. In non-parametric method training data is used for classification.Parametric method uses some parameters of handwritten samples for there classification. Generally, parametric methods are preferred than non-parametric method.

Bellegarda et al. ,1994[26] has proposed an algorithm which is based on the production of the feature vectors which reprsents each character in one or more feature spaces by using the mixture modeling technique and gaussian means clustering.

Veltman et al.,1994[27] has proposed a online handwritten character recognition system. which works on the basis of HMM(Hidden Markov Model).This system isolates online handwritten characters and it achievs 6.9% as an average error rate for handwritten character recognition.

Rigoll et al. ,1996[28] gives the comparison of continuous and discrete density Hidden Markov Modelfor the recognition of cursive handwriting. It is observed that the discrete density model gives better results as compared to the continuous models. But it is reverse in the case speech recognition.

Shimodiara et al. ,2003[29] has presented an interface for the recognition of hand writing. In this interface users have to write the characters continouosly without giving the space or pause on a small text box. This system uses the Hidden Markov Model for the recognition of handwriting.

Funanda et al. ,2004[30] has proposed a system which uses the HMM for the recognition of the online handwritten recognition. The proposed system reduces the usage of memory and also it improved the recognition rate of online handwritten characters.

4) Elastic matching Method

In Elastic matching the similarity between two pattrens is considered and the pattren which have to be recognized is matched with the stored pattren on the basis of their similarties. It is also called as nonlinear template matching or flexible matching.

Fujimoto et al., 1976[31] used the elastic matching method for the scanned images recognition. This system works on the basis of FORTAN program which is later converted into ASCII code. The digitized program gives the direction of the pattrens in the coded form to make a sequence of coordinate points. Distance between the points is calculted by using the dynamic programming formula.

Kurtzberg ,1987[32] has proposed the procedure for recognition of the unconstrined handwritten discrete symbols by using elastic matching with a set of prototypes produced by different writers. He also proposes the analysis of the features using the elastic matching method which removes the unlikely prototypes.

Ueda et al.,1990[33] proposed a system which uses elastic matching method to match shapes after various level of refinement, discriminate arbitrary shapes and to identify shapes. This system uses the dynamic programming for matching the

inflected pointsbetween the model.

Wakahara et al,1997[34] proposed a method for recognition of characters by using distant tolerant stroke matching method which uses affine transformations of the strokes. These stroke based transformation gives best matches of each strokes with the refference pattrens. They used this method on Kanji characters for recognition and achieved 98.4% recognition rate for freely written data in square style and 96% for cursive handwriting and fast written characters.

Li et al.,1997[35] has proposed a method for online handwritten character recognition, in which both alphabets and numerics are included. The proposed method recognizes the characters on the basis of sequence of dominant points and sequence of writing directions in storkes of characters. Directional information is used for pre-classification of the characters and for fine classification positioned information is used. The fine and pre classification is done by using elastic techniques and dynamic programming.

Choi et al., 2002[36] presented an algorithm with the combination of the zernike moments and projection method for template matching. This algorithm works in two stages. In first stage, by calculating a low cost feature the matching characters are selected. In second stage,by using the zernike moments matching characters are proposed by the rotation invariant template matching method.

F. Post processing

In Post processing phase, results are corrected out which are misclassified by applying different methods and approaches like using linguistic knowledge, analysing the individual characters in the form of graphs or with help of shape recognition.

Pitrelli et al., 2002[37] has proposed different methods like estimation of posterior probability for correctness, likelihood ratios and number of scoring functions which increase the confidence scores of recognition of words or characters.

Carbonnel et al.,2004[38] has designed a system for handwritten word recognition which works on optmized lexical post processing. In which segmentation and recognition errors are corrected by using the lexical information. This lexical post processing works on two phases. In first phase, the search space is reduced by replacing the lexicons by sub-lexicons during the process of recognition. In second phase, a edit distance is used for identifying the characters with the help of the selected sub-lexicon.

Namboodiri et al. ,2007[39] has proposed a system which uses the metric information to enhance the recognition rate of classical indian poetry. The proposed algorithm is combined with other post processing methods and it provides better correcting results.

III. COMPARISON

Table I Font Sizes for Papers

Author(s)	Data set	Method	Recognition/ Error rate
Nouboud et al. (1991)	(0-9) digits and (a-z) lowercase characters, 23 symbols with writer dependent system.	Chain Code Method	96%
Matic et al. (1993)	A-Z characters and 0-9 digits, 7 symbols with writer independent systems.	Neural Network Method	2.5%
Bontempi et al.(1994)	(0-9) digits and (a-z) characters with writer independent systems.	Neural Networks	1.9%(0.8%)
Li et al.(1997)	(A-Z) characters,(0-9) digits and (a-z) characters with writer independent system.	Chain Code Method	91%
Yaeger et al. (1998)	(A-Z) characters, (0-9) digits, 23 symbols with writer independent system.	Multi Layer Perceptron	21.3% (7.2%)
Hu et al. (2000)	(a) 500, 1000 and 2000 unipen database. (b) 5000, 10000 and 20000 unipen database.	Hidden Markov Model	91.8%, 90.5% and 87.2% for (a) dataset and 83.2%, 79.8% and 76.3% for (b) dataset.
Connell et al.(2000)	Devanagiri characters.	Hidden Markov Model	86.5%
Jaeger et al. (2001)	5000, 20000 and 50000 English word dictionaries.	Multi State Time Delay Neural Networks	96%, 93.4% and 91.2%
Funanda et al. (2004)	Kanji, Katakana, Hirangana, Western alphabets and symbols with writer independent system.	Hidden Markov Model	91.34%
Joshi et al. (2005)	Devanagiri characters with writer dependent system.	Feature Based Matching	94.49%

IV. CONCLUSIONS

A review of various techniques has been presented in this paper. A systematic flow of OCR system is discussed and previous work done in this field is surveyed phase by phase. It has been seen that a tradeoff is must betwwen the accuracy of the system and in the processing time of the system. One cannot achieve both in same OCR system as fast recognition rate will results in wrong results and vice versa. So, it is very important to choose the techniques for developing an OCR syatem as per the requirement of application.

REFERENCES

[1] Beigi, H., Nathan, K., Clary, G. J., and Subhramonia, J., 1994. Size normalization in unconstrained online handwritng recognition, Proceedings ICIP, pp. 169-173.

[2] Unser, M., Aldroubi, A., Eden, M., 1993. B-Spline signal processing: part II - efficient design and applications. IEEE Transactions on Signal Processing, vol. 41, no. 2, pp. 834-848.

[3] Kavallieratou, E., Fakatakis, N., Kolkkinakis, G., 2002. An unconstrained handwriting recognition system. International Journal of Document Analysis and Recognition, vol. 4, no. 4, pp. 226-242.

[4] Uchida, S., Taira, E., Sakoe, H., 2001. Nonuniform slant correction using dynamic programming. Proceedings of International Conference on Document Analysis and Recognition. pp. 434-438.

[5] Guerfali, W and Plamondon, R, 1993. Normalizing and restoring online handwriting. Pattern Recognition, vol. 26, no. 3, pp. 419.

[6] Rocha, J. and Pavlidis, T., 1994. A shape analysis model with applications to a character recognition system. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 4, pp. 393-404.

[7] Hu, J., Lim, S. G. and Brown, M. K., 2000. Writer independent on-line handwriting recognition using an HMM approach. Pattern Recognition, vol. 33, no. 1, pp. 133-147.

[8] Hu, J., Rosenthal, A. S. and Brown, M. K., 1997. Combining high level features with Sequential local features for online handwriting recognition. Proceedings of Italian Image Process conference, pp. 647-657.

[9] Verma, B., Blumenstein, M., Ghosh, M., 2004. A novel approach for structural feature extraction: Contour vs. direction. Pattern Recognition Letters, vol. 25, no. 9, pp. 975-988.

[10] Joshi, N., Sita, G., Ramakrishnan, A. G., Deepu, V., and Madhvanath, S. 2005. Machine recognition of online handwritten Devanagari characters. Proceedings of International Conference of Document Analysis and Recognition, pp. 1156-1160.

[11] Schplachbach, A. and Bunke, H., 2007. Fusing asynchronous feature streams for on-line writer identification. Proceedings of International Conference on Document Analysis and Recognition, pp. 1-5.

[12] Tappert, C. C., Suen, C. Y., Wakahara, T., 1990. The state of the art in online handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 8, pp. 787-808.

[13] Yanikoglu, B. and Sandon, P. A., 1998. Segmentation of off-line cursive handwriting using linear programming. Pattern Recognition, vol. 31, pp. 1825-1833.

[14] Blumenstein, M. and Verma, B., 1999. A new segmentation algorithm for handwritten word recognition. Proceedings of the International Joint Conference on Neural Networks, pp. 878-882.

[15] Plamondon, R. and Srihari, S. N., 2000. Online and offline handwriting recognition: A comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84.

[16] Matic, N., Guyon, I. and Vapnik, V., 1993. Writer-adaptation for online handwritten character recognition. Proceedings of International Conference on Pattern Recognition and Document Analysis, pp. 187-191.

[17] Cho, S., 1997. Neural-network classifiers for recognizing totally unconstrained handwritten numerals. IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 43-53.

[18] Kimura, Y., Wakahara, T. and Odaka, K., 1997. Combining statistical pattern recognition approach with neural networks for recognition of large-set categories. Proceedings of International Conference on Neural Networks, vol. 3, pp. 14291432.

[19] Yaeger, L. S., Webb, B. J. and Lyon, R. F., 1998. Combining neural networks and context-driven search for online, printed handwriting recognition in the NEWTON. AAAI's AI Magazine, vol. 19, no. 1, pp. 73-89.

[20] Jaeger, S., Manke, S., Reichert, J., Waibel A., 2001. Online handwriting recognition: the Npen++ Recognizer. International Journal of Document Analysis and Recognition, vol. 3, no. 3, pp. 169-180.

[21] Shintani, H., Akutagawa, M., Nagashino, H., Kinouchi, Y., 2005. Recognition mechanism of a neural network for character recognition. Proceedings of Engineering in Medicine and Biology 27th Annual Conference, pp. 6540-6543.

[22] Connell, S. D., Sinha, R. M. K. and Jain, A. K., 2000. Recognition of unconstrained online Devanagari characters. Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 368-371.

[23] Nouboud, F. and Plamondon, R., 1991. A structural approach to online character recognition: System design and applications. International Journal of Pattern Recognition and Artificial Intelligence, vol. 5, no. 1/2, pp. 311-335.

[24] Bontempi, B. and Marcelli, A., 1994. A genetic learning system for on-line character recognition. Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 83-87.

[25] Duneau, L. and Dorizzi, B., 1994. Online cursive script recognition: a system that adapts to an unknown user. Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 24-28.

[26] Bellegarda, E. J., Bellegarda, J. R., Nahamoo, D. and Nathan, K. S., 1994. A fast statistical mixture algorithm for online handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 12, pp. 1227-1233.

[27] Veltman, S. R. and Prasad, R., 1994. Hidden markov models applied to online handwritten isolated character recognition. IEEE Transactions on Image Processing, vol. 3, no. 3, pp. 314-318.

[28] Rigoll, G., Kosmala, A., Rottland, J. and Neukirchen, C., 1996. A comparison between continuous and discrete density hidden Markov models for cursive handwriting recognition. Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 205-209.

[29] Shimodaira, H., Sudo, T., Nakai, M., Sagayama, S., 2003. Online overlaidhandwriting recognition based on substroke HMMs. Proceedings of International Conference on Document Analysis and Recognition, pp. 1043-1047.

[30] Funanda, A., Muramatsu, D., Matsumoto, T., 2004. The reduction of memory and the improvement of recognition rate for HMM on-line handwriting recognition. Proceedings of IWFHR, pp. 383-388.

[31] Fujimoto, Y., Kadota, S., Hayashi, S. and Yamamoto, M., 1976. Recognition of hand printed characters by nonlinear elastic matching. Proceedings of the Third Joint Conference on Pattern Recognition, pp. 113-118.

[32] Kurtzberg, J. M., 1987. Feature analysis for symbol recognition by elastic matching. IBM Journal of Research and Development, vol. 31, no. 1, pp. 91-95.

[33] Ueda, N. and Suzuki, S., 1990. Automatic shape model acquisition using multiscale segment matching. Proceedings of International Conference on Pattern Recognition, pp. 897-902.

[34] Wakahara, T. and Odaka, K., 1997. Online cursive kanji character recognition using stroke-based affline transformation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 12, pp. 1381-1385.

[35] Li, X. and Yeung, D. Y., 1997. Online handwritten alphanumeric character recognition using dominant points in strokes. Pattern Recognition, vol. 30, no. 1, pp. 31-44.

[36] Choi, M. and Kim, W., 2002. A novel two stage template matching method for rotation and illumination invariance. Pattern Recognition, vol. 35, no. 1, pp. 119129.

[37] Pitrelli, J.F. and Perrone, M.P., 2002. Confidence modeling for verification postprocessing for handwriting recognition. Proceedings of IWFHR, pp. 30-35.

[38] Carbonnel, S. and Anquetil, E., 2004. Lexicon organization and string edit distance learning for lexical post-processing in handwriting recognition. Proceedings of IWFHR, pp. 462-467.

[39] Namboodiri, A. M., Narayanan, P. J. and Jawahar, C. V., 2007. On using classical poetry structure for Indian language post-processing. Proceedings of International Conference on Document Analysis and Recognition, pp. 1238-1242.