ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Handwritten Document Editor: A Review

Sumit Nalawade[1], Rashmi Welekar[2], Raj Dugar[3]

M.Tech. Student, Department of Computer Science & Engineering, Shri Ramdeobaba College Of Engineering & Management, Nagpur, India [1]

Assistant Professor, Department Of Computer Science & Engineering, Shri Ramdeobaba College Of Engineering & Management, Nagpur, India [2]

Module Lead, Persistent Systems Limited, Nagpur, India [3]

ABSTRACT: In the current era, handwriting recognition & personalization of document is becoming concern of many researchers. It has most significant applications in many fields such as Optical Character Recognition, Security Systems etc. Also in daily uses such as banks, post offices, businesses. Personalization is key to the future of one's identity. This paper provides a comprehensive review of various methods used for character recognition such as Neural Networks, Support Vector machines, K-NN classifier and provides a short review of Binary Coded Genetic Algorithm which can be used for speed optimization of the classification process. This paper also gives a snapshot of Marching Squares Algorithm which is a basic procedure for tracing contour points on images.

KEYWORDS: OCR, K-NN classifier, Genetic Algorithm, Marching Squares Algorithm, Binary Gradient Vectors, Image Processing.

I. INTRODUCTION

Character recognition process comprised of extracting the features of script and individual characters and identification of characters based of features. The difficulties faced by these systems are variations in writing styles, heights & widths of characters, slanting angles. The major technique used for offline character recognition are neural network, support vector machines. Some authors have proposed systems which employed hidden markov model for the same. The problem with these approaches increase with increasing size of training data. To solve this problem we need an optimization technique which will decreased the running time of the classifier. Binary coded genetic algorithm along with K-NN classifier can be used for clustering to solve this optimization problem.

Yonghong Song [1], gives details about multi-language document images where many different types of characters need to be recognised. His document consists different languages such as Chinese, Japanese, English and also their combination. It uses genetic algorithm for feature fusion and patch type classification. At last uses markov random field model for post processing. Rahul Kala [2] in his works deals with graph theory and studies various fonts generated by graph. He uses different styles in which characters can be written and uses offline handwritten recognition for identifying characters by converting them into graphs. Finally implemented the genetic algorithm to optimize the results. S. Impedovo and F. M. Mangini [3] proposed a novel technique for handwritten digit classification which uses binary coded genetic algorithm for optimization and speed up of the K-NN classifier. The main idea behind using genetic algorithm is to reduce number of comparisons by forming cluster centers. Vijay Patil and Sanjay Shimpi [4] uses neural network for character recognition in which he created character matrix and key is suitable network structure. He also uses back propagation algorithm to calculate errors and modifying weights.

The English script consists of total 94 different shapes which includes capital letters, small letters, digits and standard special symbols. Fig. 1 shows a handwritten sample of this complete standard character set. Many systems or methods are available to recognize them separately but building a recognition system for all these shapes together is a difficult task. We need to identify the unique features of individual shapes and used proper dissimilarity measures to distinguish among them. There can be differences in height, size and width of characters. Some characters may have similar looking such as 1 (DIGIT ONE) and l (LATIN SMALL LETTER L). Different sets of characters are formed which have different look due to different environment of writing.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Fig. 1. Set of standard characters in English language.

The editor system we are trying to put to work is a twofold approach which involves character recognition and handwritten font generation. In this paper we are reviewing methods for the two and will comment on which can be the best suitable approach for our system to be built.

II. RECOGNITION METHODS OVERVIEW

This section deals with different approaches used to recognize the characters from documents. This recognition methods have genetic algorithm along with K-NN classifiers, neural networks.

Character extraction from handwritten can be done at three different steps: on the level of text [5], then word [6], [7], [8],[9] and at last characters [10], [11], [12]. Methods which are performed at text line level are not difficult to implement and can be implemented for different language, but can be done only for specified formats. At character level, there are difficulty when character appear to be curved and overlapping. Considering the limitation of such factors the author deals only with the word level of documents. The following figure 2 shows the details of the flowchart of handwritten character extraction.

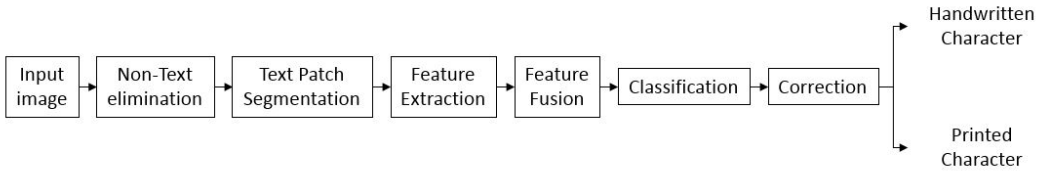

Fig. 2. Flowchart of character handwritten extraction

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

The input is the text image, from this text image some non-significant need to be eliminated. After this text patch segmentation algorithm needs to be implemented from which feature extraction are extracted and then the fusion of the feature vector is formed which is then provided for classification to which some correction needs to be done.

Previous text patch segmentation method utilizes simple operation such as mathematical morphological operation and simple neighborhood rules.

The handwritten recognition mainly involved three steps such as segmentation, pre-processing and recognition. Segmentation means breaking of the lines, word and lastly getting all the characters separated. This step involves identifying the boundaries of the character and separating them for further processing. Pre-processing of image can be used as initial steps for the recognition system. Once the input image is available in unique condition, it may be processed for recognition. The generation of graph algorithm takes image as input and return the graph of the image. The whole procedure requires the principle of graph theory and coordinate geometry. Figure 3 shows the details of the graph theory algorithm.

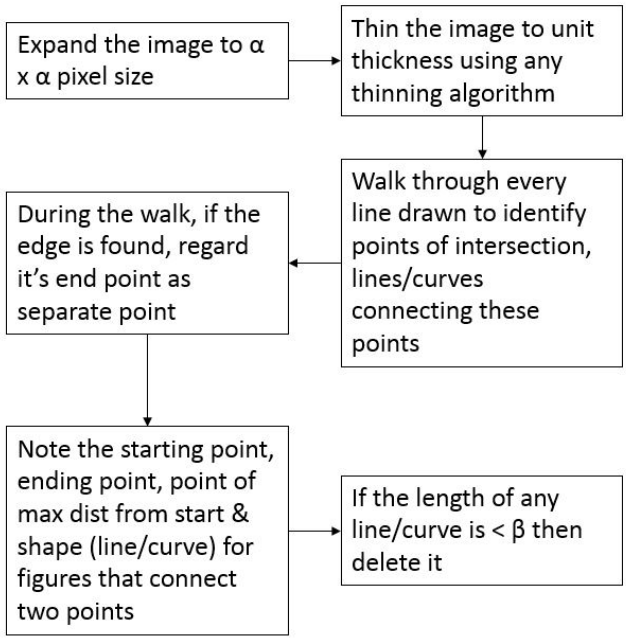

Fig. 3. Details of graph theory algorithm

There is need to form the matrixes of each character in such a way to understand how to take out the binary input image from the matrix and how to manipulate the binary output image which the computer interprets.

S. Impedovo presents a new clustering approach for improving handwritten digit classification. The K-NN classifier [15] is used with a binary coded genetic algorithm [13]. Forming the clusters is considered as optimization problem in this approach. The training data is drastically reduced to only a handful of cluster centers then K-NN classifier will use pre identified clusters to classify digits.

Feature extraction is an important task in any classification technique. This approach extracts a binary gradient vector which will hold information about the shape of characters. Each shape image is applied with a grid of different resolutions [17] for fine grained details as well as overall shape details. Inside each cell Sobel operator is applied to get information about the gradient directions. Eight directional chaincodes are used to represent direction counts. The values resulted in each cell are aggregated over a histogram and a threshold [16] is applied to get an 8-bit gradient vector. Together all cells at different resolutions form a 392-bit binary feature vector.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

This approach uses Sokal-Michener dissimilarity measure [14] to differentiate and analyze similarity between the two binary vectors.

To find the clusters some predefined threshold is decided. If the input training sample yields in dissimilarity less than this threshold then the sample is added to the cluster and center is refined, else the sample forms a new cluster center for the further training samples. Genetic algorithm is applied for finding best cluster solutions. A single point crossover is applied to fixed number of individuals and the generated offspring population is tested for fitness test. Passing candidate are taken for next iteration and mutation. This approach reduced 6000 comparisons to 701 which is very significant improvement in speed.

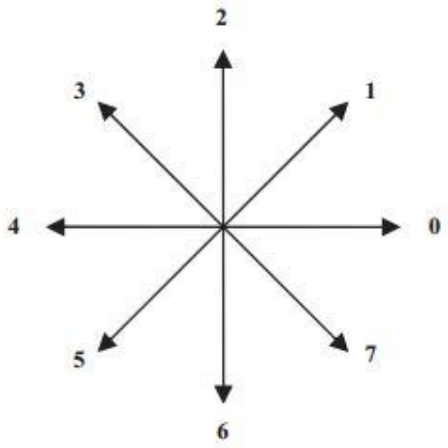

Fig. 4. Eight direction of chaincodes

III. IMAGE CONTOURING TECHNIQUES

This section reviews few available image contouring techniques which are helpful in tracing contour points of handwritten characters for creating a font of one's handwriting.

Marching Squares Algorithm [18] is one of the popular techniques used for image contour tracing. This is direct descendent of 3D marching cubes algorithm [19] that generates contours for a 2D scalar field. This algorithm considers a 2x2 grid for processing each cell independently. The start point of the algorithm is any point on the edge of the image under consideration and the stopping condition is satisfied when the marching square reaches where it started.

The procedure to marching squares is pretty simple. A 2x2 grid mask is applied to point on edge and cell index is calculated for particular point. For standard algorithm consideration this value varies from 1 to 15. Based on this index value the direction of the square to march is decided by looking into a prebuilt lookup table. The grid or the square is then advanced into determined direction pushing current point to the list of contour points. Similar procedure is repeated unless starting point is revisited.

Fig. 5. 2x2 grid for marching squares cell index calculation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

In another approach A. M. Baumberg& D. C. Hogg present a contour tracking techniques using active shape model. [20] The method demonstrate a combination of shape model or the point distribution model with dynamic filtering to track a non-rigid body in motion. A training shape of fixed number of control points is aligned to reference shape to analyze the deviations. The projection of the shape in model is obtained by translation, rotation & scaling of the model frame. The curve is sampled at fixed regular intervals for convenience. At each new frame search for the sample points is made along normal for measurements. The retained observed feature is nothing but the point of maximum contrast.

The shape part of model vector is modelled as a simple discrete stochastic process. Then a tracking filter is applied for decoupling shape, alignment & translation effects in the model from previous operations. Updating state estimate of a system with discrete measurement is the work left for Kalman Filter. Then the application of alignment filter, shape filter and the global shape constraint provides a feasible contour shape.

IV. CONCLUSION AND FUTURE WORK

This paper gives a brief review about the recognition systems and the contouring techniques which are the heart of our proposed editor system. Since our editor system prefers speed over accuracy the method with genetic clustering should give promising result. In accordance to the second step the marching squares algorithm provides more simplicity and speed for contouring and it's a widely used and refined method with over average accuracy. The combo of the two should allow us two build our system as far as primary objective is of concern. Later refinements can be applied with the use of more advanced techniques.

REFERENCES.

1. Yonghong Song, Guilin Xiao, Yuanlin Zhang, Lei Yang, Liuliu Zhao, "A Handwritten Character Extraction Algorithm for Multi-language Document Image", International Conference on Document Analysis and Recognition 2011.
2. Rahul Kala, Harsh Vazirani, Anupam Shukla, Ritu Tiwari, "Offline Handwritten Recognition using Genetic Algorithm", IJCSI Vol. 7, Issue. 2, March 2010.
3. S. Impedovo, F. M. Mangini, "A Novel Technique For Handwritten Digit Classification using Genetic Clustering", International Conference on Frontiers in Handwriting Recognition, 2012.
4. Vijay Patil, Sanjay Shimpi, "Handwritten English character recognition using neural network", Elixir International Journal, 2011.
5. Srihari SN, Shim YC, Ramanprasad V. "A System to Read Names and Address on Tax Forms", CEDAR, SUNY, Buffalo, N.Y., 1994.
6. Guo JK, Ma MY.,"Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models", Proc. Int'l Conf. Document Analysis and Recognition, pp:439-443, 2001.
7. Zheng Y, Li H, Doermann D.,"Machine Printed Text and Handwriting Identification in Noisy Document Images", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,issue 26, 2004.
8. Shetty S, Srinivasan H, Beal M, et al. "Segmentation and labeling of documents using Conditional Random Fields", Proceedings of SPIE, 2007.
9. Peng X, Setlur S, Govindaraju V, et al. "Markov Random Field Based Text Identification from Annotated Machine Printed Documents", 10th International Conference on Document Analysis and Recognition, pp431-435, 2009.
10. Kuhnke K, Simoncini L, Kovacs-V ZM.,"A System for Machine-Written and Hand-Written Character Distinction", Proc. Int'l Conf. Document Analysis and Recognition, pp:811-814, 1995.
11. Zheng Y, Liu C, Ding X. "Single Character Type Identification", Proc. SPIE Conf. Document Recognition and Retrieval, pp:49-56, 2002.
12. Koyama J, Kato M, Hirose.,"A Handwritten Character Distinction Method Inspired by Human Vision Mechanism", Lecture Notes In Computer Science, 2008.
13. D. Beasley, D.R. Bull, R.R. Martin, "An Overview of Genetic Algorithms: Part 1, Fundamentals", University Comput. ,Vol 15,n.2,pp. 58-69, 1993.
14. R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships", University of Kansas Scientific Bulletin 38, pp. 1409-1438, 1958.
15. H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition", in Proc.CVPR, 2006.
16. N. Otsu. "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man, and Cybernetics, 9(1), pp: 62-66 1979.
17. S. Impedovo, G. Pirlo, R. Modugno, A. Ferrante, "Zoning Methods for Handwritten Character Recognition: An Overview", Proc. 12th ICFHR, Kolkata (India), 16-18 Nov., pp. 329-334, 2010.
18. Maple, C. "Geometric design and space planning using the marching squares and marching cube algorithms", Proc. 2003 International Conference Geometric Modeling and Graphics: 90–95.
19. Chien-Chang Ho, Fu-Che Wu, Bing-Yu Chen, Yung-Yu Chuang, Ming Ouhyoung, "Cubical Marching Squares: Adaptive Feature Preserving Surface Extraction from Volume Data"EUROGRAPHICS (Guest Edition) Volume 24, Number 3, 2005.
20. A. M. Baumberg& D. C. Hogg, "An Efficient Method for Contour Tracking using Active Shape Models", University of Leeds, School of Computer Studies Research Report Series, Report 94.11, 1994.