Evaluation of Feature Extraction and Classification Techniques on Special Symbols

Sanjay S. Gharde, Vidya A. Nemade, K. P. Adhiya

Abstract— The Symbol recognition is a significant area in computer vision that involves the recognition of symbols in an image or video. Symbol recognition is used in a large number of different applications like converting the text document into other electronic document formats, forming mathematical expression, maps, musical notations etc. When someone copies the data from any text document like any PDF file instead of that original characters or special symbols, garbage value would be copied at some places. So identification of special symbols requires higher accuracy at the time of recognition. This paper describes the analytical study of various techniques which will be useful for implementation of the system. Some existing techniques of feature extraction, classification and recognition of symbols and characters are compared in this paper.

Index Terms— Classification, feature extraction, preprocessing, segmentation, symbol recognition, support vector machine,

1 Introduction

Symbol recognition is used in a large number of diverse applications such as:

- Interpreting and converting scanned engineering drawings and circuit diagrams into other electronic manuscript formats.
- Recognizing and locating trademarked content.
- Querying images from databases based on shape and
- Recognizing characters and words within an electronic document [1].

The exsisting symbol recognition approaches can be divided usually into two groups: statistical approaches and structural approaches. Normally structural approaches are vector-based in which symbol is decomposed into some vector-based primitives at first such as dominant points, lines and arcs. As for statistical approaches, the pixel is typically the primitive. The most important stastical descriptors include plain binary image, moment invariants, ring projection and shape context [2]. In many applications it is necessary to copy the contents from some original documents which may be in PDF or any other format. So, while copying the data from document, when it encounters a special symbol it remains unread in the copied document, or instead of that special symbol random value placed in the copied document. So it is very difficult to read the original document. While reading the printed English text if a special symbol found then main task is to identify that particular special symbol. After recognition of that symbol same process continues i.e. if again symbol found while reading the rest of the document processing will done on it. Processing of symbol consists of following steps -Preprocessing, Segmentation, Feature extraction, Classification and Recognition.

The architecture of symbol recognition consists of general image processing steps as shown in Fig. 1. Usually image consisting of symbol or character is acquired through scanner or any other digital device. Preprocessing is the next step to produce the clean image after image acquisition and to make it suitable for segmentation and feature extraction. Finally support vector machine or any other suitable classifier can be used for recognition purpose.

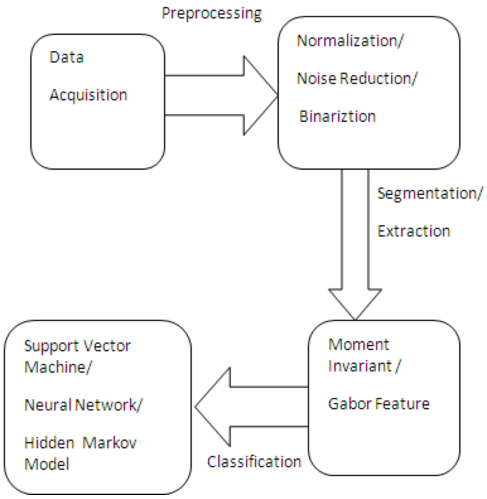
Preprocessing
Data Acquisition
Normalization/ Noise Reduction/ Binariztion
Segmentation/ Extraction
Moment Invariant / Gabor Feature
Classification
Support Vector Machine/ Neural Network/ Hidden Markov Model

Fig. 1. Architecture of Symbol Recognition

2 Related Work

The history of character recognition can be traced as early as 1900, when the Russian scientist Tyuring attempted to develop an assist for the visually handicapped. The first character recognizers developed in the middle of the 1940s with the

- *Sanjay S. Gharde is currently working as assistant professor in Computer department of SSBT College of Engineering and Technology, Jalgaon. PH-+91-9422344964. INDIA, E-mail: sanjay_gharde358@yahoo.com*
- *Vidya a. Nemade is currently pursuing masters degree in computer science engieeringin from North Maharashtra University, Jalgaon, INDIA, PH-+91-9561473161. E-mail: vidya.nemade@gmail.com*
- *K. P Adhiya is currently Associate Professor in Computer Engineering Department of SSBT College of Engineering and Technology, Jalgaon, INDIA. PH-+91-9421516499. E-mail: kpadhiya@yahoo.com*

growth of digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine-printed text or upon a small set of handwritten text or symbols. Machine-printed Character recognition systems normally used template matching. For handwritten text, low-level image processing techniques have been used on the binary image to extract feature vectors, which are then fed to statistical classifiers.

Successful, but inhibited algorithms have been implemented mostly for Latin characters and numerals. However, some studies on Japanese, Chinese, Hebrew, Indian, Cyrillic, Greek, and Arabic characters and numerals in both machine-printed and handwritten cases were also initiated. The commercial character recognizers were available in the 1950s. Historical review of charcter recognition research and growth during this period can be found in and for off-line and on-line cases, respectively. In the early 1990s, image processing and pattern recognition techniques were efficiently united with artificial intelligence (AI) methodologies. Nowadays, in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras, and electronic tablets, we have efficient, modern use of methodologies such as neural networks (NNs), hidden Markov models (HMMs), fuzzy set reasoning, and natural language processing. [3]

3 Proess Of Symbol Recognition

The symbol recognition normally includes the following parts: Preprocessing, Feature extraction, and Classification.

3.1 Preprocessing

Preprocessing of a character or symbol is important because it affect the recognition rate very much. Preprocessing brings the raw data into some specific format and it gives good results. The main objectives of preprocessing are [3]:
1) Noise reduction
2) Normalization
3) Compression. Some of the exsisting preprocessing techniques are given in TABLE 1. Recognition system consists of following modules:

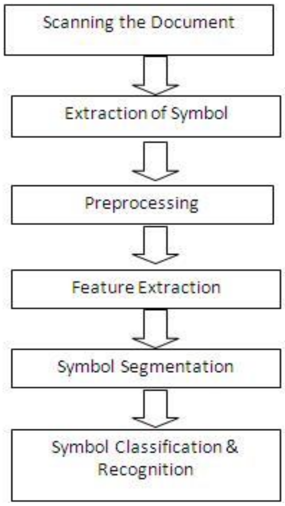

Fig. 2. Process of Symbol Recognition

There are several preprocessing techniques available to produce the clean image. It consists of many steps like Filtering, Morphological operation, Normalization, Thinning, Resampling.

3.1.1 Noise Reduction-

The noise, introduced by the optical scanning device or the writing instrument, causes disconnected line segments and gaps in lines, filled loops, etc. Hundreds of available noise reduction techniques can be categorized in major groups. Some of the exsisting preprocessing techniques are as follows:

TABLE 1
Exsisting preprocessing Techniques

Author	Preprocessing Techniques
Vnaita Mane, Leena Ragha[4]	Filtering, Morphological Operations, Normalization
Birendra Keshari, Stephen Watt[5]	Gaussian Smoothing, Resampling
Gong Xin, LI Cuiyun, PEI Jihong, XIE Weixin[6]	Noise Removal, Smoothing, Data point Rearrangement
Dewi Nasien, Habibollah Haron, Siti Yuhaniz[7]	Thinning
Shubhangi D. C, Prof. P.S.Hiremath[8]	Normalization, Thinning
Lie Hu, Richard Zanibbi[9]	Duplicate Point Filtering, Size Normalization, Smoothing, Resampling
J.Pradeep, E. Srinivasan, S.Himavathi, [10]	Binarization, dilation
T. Cheng, J. Khan, H. Liu and D. Y. Y. Yun[11]	autocropping, centralization, and normalization
Birendra Keshari, Stephen Watt[12]	Smoothing, Filling intermediate points, resampling, Size normalization
Widad Jakjoud , Azzeddine Lazrek[14]	Filtering (Gaussian filter) Binarization

3.1.1.1 Filtering:

This aims to remove noise and diminish false points, usually introduced by irregular writing surface and/or poor sampling rate of the data acquisition device. Filters can be planned for smoothing, sharpening, thresholding, removing slightly textured or colored background, and contrast adjustment purposes. [3]

3.1.1.2 Morphological Operations:

The basic idea behind the morphological operations is to filter the document image replacing the convolution operation by the logical operations [3]. Various morphological operations such as bridge, fill, clean, majority and thin can be intended to connect the broken strokes, decompose the connected strokes, smooth the contours, thin the characters, and extract the boundaries. Therefore, morphological operations can be successfully used to remove the noise on the document images due to low quality of paper and ink, as well as irregular hand movement [4].

3.1.2 Normalization:

Normalization methods try to remove the variations of the writing and acquire standardized data. The following are the basic methods for normalization.

3.1.2.1 Skew Normalization

Due to inaccuracies in the scanning process and writing style, the writing may be slightly tilted or curved within the image. This can harm the effectiveness of later processing and, theefore, should be detected and corrected (e.g., “9” and “g”) [3].

3.1.2.2 Size Normalization:

This is used to adjust the character size to a certain standard. Methods of character recognition may apply both horizontal and vertical size normalizations [3].

3.1.2.3 Slant Normalization:

Slant normalization is used to normalize all characters to a standard form. The most common method for slant estimation is the calculation of the average angle of near-vertical elements [3].

3.1.3 Compression:

Compression is mainly used to reduce the storage requirement for image. Two popular compression techniques are thresholding and thinning.

3.1.3.1 Thresholding (Binarization)

In order to reduce storage requirements and to increase processing speed, it is often desirable to represent gray-scale or color images as binary images by picking a threshold value. Two categories of thresholding exist: global and local. Global thresholding picks one threshold value for the entire document image and local thresholding uses the different values for different regions [3].

3.1.3.2 Thinning (Skeletonisation)

This process extracts the shape information of characters or symbols. Skeletonisation is also called thinning. Skeletonisation refers to the process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and this process also reduces the processing time [3].

From Table 1 it is observed that Filtering, Normalization, Binarization can be used as a preprocessing techniques.

3.2 Evaluation of Feature Extraction Techniques

Feature extraction is the name given to a family of procedures

TABLE 2
Review of Exsisting Feature Extraction Methods

Author	No. of Samples	Feature Extraction Techniques
Birendra Keshari, Stephen Watt[5]	Dataset with 48 symbols	Co-ordinates of resampled points, sine & cosines of the angle made by segments and turning angle
Gong Xin, LI Cuiyun, PEI Jihong, XIE Weixin[6]	Graphic library contain 110 symbols	Normalizes distance measure, rotation angle of the strokes
Dewi Nasien, Habibollah Haron, Siti Yuhaniz[7]	Lowercase-189,411 samples, Uppercase-217,812 samples	Freeman chain code (4-neighbourhood)
Shubhangi D. C, P.S. Hiremath[8]	26000 Samples of handwritten english character, 5000 samples english handwritten digits	Height, Width,Slantness, Average pseudo pressure, number of Strokes etc.
Lie Hu, Richard Zanibbi [9]	20281 samples in training set, 2202 samples in testing set	Pen up / down, speed, normalized x & y co-ordinate, cosine & sine of curvature,vicinity aspect etc.
J.Pradeep, E. Srinivasan, S.Himavathi [10]	50 dataset	Without any feature Extraction
T. Cheng, J. Khan, H. Liu and D. Y. Y. Yun[11]	34 symbols	Movement invariant
Birendra Keshari, Stephen Watt[12]	137 unique Mathematical symbols	Intersections, loops, Height-weidth ratio, Number of strokes, Initial / End position etc.
Xue Dong Tian, Hai-Yan Li, Xin-Fu Li and Li-Ping Zhang[13]	100 Chinese mathematical literature, 3000 mathematical formula	Gabor filter, Parameter selection (standard deviation,wavelength), Elastic meshing
Widad Jakjoud , Azzeddine Lazrek[14]	Symbol & error database	Countour & region approaches
Shailedra Shrivastava, Sanjay S. Gharde[15]	2000 samples	Moment Invariant and Affine Moment Invariant

for measuring the relevant shape information contained in a pattern so that the task of classifying the pattern becomes easy. The feature extraction stage analyses a text segment and selects a set of features that can be used to uniquely identify the text segment.

The task of human expert is to select features that permit effective and efficient recognition of pattern. Feature extraction is a very significant in recognition system because it is used by the classifier to classify the data. This can be achieved by some of the following:

3.2.1 Chain Code

Chain code is demonstration technique which is useful for image processing, shape analysis and pattern recognition fields. Chain code representation provides the boundary of character image in which the codes represent the direction of where is the location of the next pixel. One such method is Freeman Chain Code having two directions of chain code, namely 4- neighborhood and 8-neighborhood [7].

3.2.2 Moment Invariant: The moment invariants (MIs) are used to evaluate seven distributed parameters of a numeral image [15].

3.2.3 Gabor Filter: 2-D Gabor filter is a somewhat complex sinusoidally modulated Gaussian function with the response in spatial domain and spatial frequency domain.

From Table 2 it is observed that moment Invariant can also be good feature extraction technique because it can be applied to different size images and this approach is invariant to rotation, scaling, and translation and can also be used as aircraft identification**,** and character recognition. Gabor filter can also be used as feature extraction because along with this Xue-Dong Tian [13] achieved high recognition rate.

3.3 Evaluation of Classification Techniques

During classification stage character or symbol is placed in the appropriate class to which it belongs. Various classifiers can be used for the recognition purpose. Some of them are: Support Vector Machine (SVM), Neural Network (NN), Hidden Markov Model (HMM)

3.3.1 Support Vector Machine

Generally, Support Vector Machines (SVM) is used for classification in pattern recognition. Support Vector machine is one of the supervised learning technique. First realistic implementation of SVM had been executed in early nineties. This strategy is introduced by Vapnik and co-workers. Support vector machine is one of the paramount techniques used for linear and nonlinear classification. The SVM classifier was formerly developed for two-class or binary classification and the challenging applications of pattern recognition led to the design of multi-class SVM classifiers using the binary SVM classifiers.

SVMs were developed to solve the classification problem, but recently they have been unlimited to solve regression problems [15].

3.3.2 Hidden Markov Model

The research work of Hidden Markov Model can be traced back to 1960’s, but the HMM is not generally used in the pattern recognition field until 1980's. This model is primarily applied in the continuous voice recognition and great success has been gained. In recent years, HMM is also be used in the handwriting recognition and cursive script recognition.

Hidden markov model is a dual random process each comes from a markov process. One of these random process is not obvious, its characteristic can only he described by the other random process's surveillance. This hidden process is a finite state process [6].

3.3.3 Neural Network

Feedforward neural networks, including multilayer perceptron (MLP), radial basis function (RBF) network, the probabilistic neural network (PNN), higher-order neural complex (HONN), etc., have been broadly applied to pattern recognition. The connecting weights are usually adjusted to minimize the squared error on training samples in supervised learning. A network using local connection and shared weights, called convolutional neural network, has reported huge success in character recognition. Thers is a Zernike moment feature based approach for Devnagari handwritten character recognition. They used an artificial neural network for classification.

TABLE 3
Evaluation of Classification method with their Recognition Rate

Author	Classifier	Rec. Rate
Gong Xin, LI Cuiyun, PEI Jihong, XIE Weixin[6]	Hidden Markov Model	85.%
Shubhangi D. C, Prof. P.S.Hiremath[8]	Multi class SVM Classifier	99.91%
Lie Hu, Richard Zanibbi[9]	Hidden Markov Model	92.31%
J.Pradeep, E. Srinivasan, S.Himavathi, [10]	Feed forward backpropogation Neural Network	90.19%
Xue-Dong Tian, Hai-Yan Li, Xin-Fu Li and Li-Ping Zhang[13]	Minimum distance Classifier	97.75%
Shailedra Shrivastava, Sanjay S. Gharde[15]	Support Vector Machine	99.48%
B.Q.Huang, M.T. Kechadi[16]	Hidden Markov Model & Multilayer Perceptron	93.43% (Avg.)
C. Pirlo, D. Impedovo[17]	Zoing Based Classification	95%
M. Hanmandlua, K.R. Murali Mohanb, Sourav Chakrabortyc, Sumeer Goyald, D. Roy Choudhurye	Neural Network, Fuzzy Logic	93.84% (NN) 98.25% (Fuzzy Logic)

From Table 3 it is observed that Gong [6] obtained a recognition rate of 85% using hidden markov model and Haung [16] obtained the recognition rate of 93.43%. Support vector machine [8], [15] provides the recognition rate upto 99.91% to 99.48%. Using minimum distance classifier Xue-Dong Tian [13] obtained the recognition rate 95.75%. C. Pirlo [17] proved that Zoing Based Classification provides the recognition rate of 95%. So it is recognized that recognition rate can be higher using support vector machine.

4 Conclusion

There are various approaches for recognition of symbols as neural network, support vector machine and hidden markov model etc. After review study it is observed that support vector machine can provides better recognition rate as compared to other classification techniques.

From the evaluation of feature extraction and classificationn techniques, system can be implemented for recognition of special symbol using moment invariant as a feature extraction technique and support vector machine as a classifier.

References

[1] Alexander Wong and William Bishop, Robust Hough-Based Symbol Recognition Using Knowledge-Based Hierarchical Neural Networks.

[2] Su Yang, “Symbol Recognition via Statistical Integration of Pixel-Level Constraint Histograms: a New Descriptor” *IEEE Transaction on Pattern Analysis and Machine Intelligence,* Vol 27, No. 2, 278-281, Feb 2005.

[3] Nafiz Arica and Fatos T. Yarman-Vural, “An Overview of Character Recognition Focused on of Character Recognition Focused on Off-Line Handwriting” 2001 IEEE.

[4] Vanita Mane, Leena Ragha, “Handwritten Character Recognition using Elastic Matching and PCA” International Conference on Advances in Computing, Communication and Control (ICAC3’09) 2009 ACM ,410-415, 978-1-60558-351-8.

[5] Birendra Keshari and Stephen Watt, “Online Mathematical Symabol Recognition using SVM’s with Features from Functional Approximation

[6] Gong Xin, LI Cuiyun, PEI Jihong, XIE Weixin,” HMM Based Online Hand-Drawn Graphics Symbol Recognition” ICSP’02 Proceedings1067-1070, 0-7803-7488-6, 2002 IEEE.

[7] Dewi Nasien, Habibollah Haron, Siti Yuhaniz ,“Support Vector Machine (Svm) For English Handwritten Character Recognition”, 2010 Second International Conference on Computer Engineering and Applications,249-252, 978-0-7695-3982-9/10 2010 IEEE.

[8] Shubhangi D. C, Prof. P.S.Hiremath ,”Handwritten English Character And Digit Recognition Using Multiclass SVM Classifier And Using Structural Micro Features”, International Journal of Recent Trends in Engineering, Vol 2, No. 2, November 2009, 193-195.

[9] Lie Hu, Richard Zanibbi, “ HMM-Based Recognition of Online Hnadwritten Mathemtical symbols Using Segmental K-means Initialization and A Modified Pen-up/down Feature” 2011 International conference on Document analysis and Recognition, 457- 462, 1520-5363/11 , 2011 IEEE.

[10] J.Pradeep, E. Srinivasan, S.Himavathi, ”Neural Network based Handwritten Character Recognition system without feature extraction” International Conference on Computer, Communication and Electrical Technology ICCCET March, 2011, 40-44, 2011 IEEE

[11] T. Cheng, J. Khan, H. Liu and D. Y. Y. Yun , “A Symbol recognition system”, 0-81864960-7193, 918-920, 1993 IEEE.

[12] Birendra Keshari and Stephen M. Watt, “Hybrid Mathematical Symbol Recognition using Support Vector Machines”.

[13] Xue-Dong Tian, Hai-Yan Li, Xin-Fu Li and Li-Ping Zhang ,”An Improved Method Based On Gabor Feature For Mathematical Symbol Recognition”, *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,* 19-22 August 2007, 1678-1682, -4244-0973-X/07 2007 IEEE.

[14] Widad Jakjoud, Azzeddine Lazrek,” Segmentation method of offline Mathematical symbol”, 978-1-61284-732-0/11 2010 IEEE.

[15] Shailedra Kumar Shrivastava, Sanjay S. Gharde, “Support Vector Machine for Handwritten Devanagari Numeral Recognition” International Journal of Computer Applications (0975 – 8887), 9-14, Volume 7– No.11, October 2011.

[16] B.Q.Huang, M.T. Kechadi,” A Fast Feature Selection Model for Online Handwriting Symbol Recognition”.

[17] G. Pirlo and D. Impedovo,”Fuzzy-Zoning-Based Classification for Handwritten Characters”, *IEEE Transactions On Fuzzy Systems,* 780-785, VOL. 19, NO. 4, AUGUST 2011, 1063-6706.

[18] M. Hanmandlua, K.R. Murali Mohanb, Sourav Chakrabortyc, Sumeer Goyald,D. Roy Choudhurye, “ Unconstrained handwritten character recognition based on fuzzy logic” 0031-3203/02,603-623, Pattern Recognition 2003.