International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 1– No.5, February 2012 – www.ijais.org

Evaluation of Classification and Feature Extraction Techniques for Simple Mathematical Equations

Sanjay S. Gharde*
Assistant Professor

Baviskar Pallavi V.*
Research Scholar,

K. P. Adhiya*
Associate Professor

* Department of Computer Engineering, SSBT, s College of Engineering & Technology, Jalgaon, Maharashtra State, INDIA.

ABSTRACT

Recognition of simple mathematical equation can applied on on-line or off-line samples. This system can applicable for publicly available dataset or researchers can prepare their own dataset for training and testing samples. In particular, we try to focus on evaluation of various methods used for recognition system. Moreover, some necessary issues in simple mathematical equation recognition will be addressed in depth.

This paper discusses various steps of recognition process for simple mathematical equations. In that, pre-processing, segmentation, feature extraction, classification and recognition for mathematical symbol as well as for simple expression is described. Among the various phases applied in recognition system, features extraction and classification method may affect the overall accuracy of the system. Therefore, various techniques applied in this context are studied and comparative analysis is prepared. This evaluation study suggests better feature extraction and classification technique for improving the recognition rate of simple mathematical equation system.

General Terms

Pattern Recognition, Mathematical symbols.

Keywords

Mathematical equation recognition; symbol recognition; support vector machine; segmentation; classification; feature extraction.

1. INTRODUCTION

Automatic recognition of printed mathematical symbols is a fundamental problem for recognition of mathematical expressions [1]. Mathematical recognition is an important problem about pattern recognition, because mathematical expression is an essential part of scientific literature and engineering discipline [2]. The input for this system is simple mathematical equations or symbols. The input of mathematical expression into computers is often more difficult than plain text, because mathematical expression is collection of various special symbols, Latin/Arabic/Greek letters, operators and English letters, digits. Mathematical symbol recognition can be done by off-line and on-line. Here the term symbol means not only basic math symbols (e.g.”*”) but simple characters (e.g. “X”) are useful in many in-line math formulae are composed of single character, denote as *“x”* in formulae “The variable *x* denotes….”

Recognition of simple equation of typeset or dataset is difficult because following reasons [1]: variable font of symbols, different size and writing styles of same equation, quality of capture image, relative position or link of symbols.

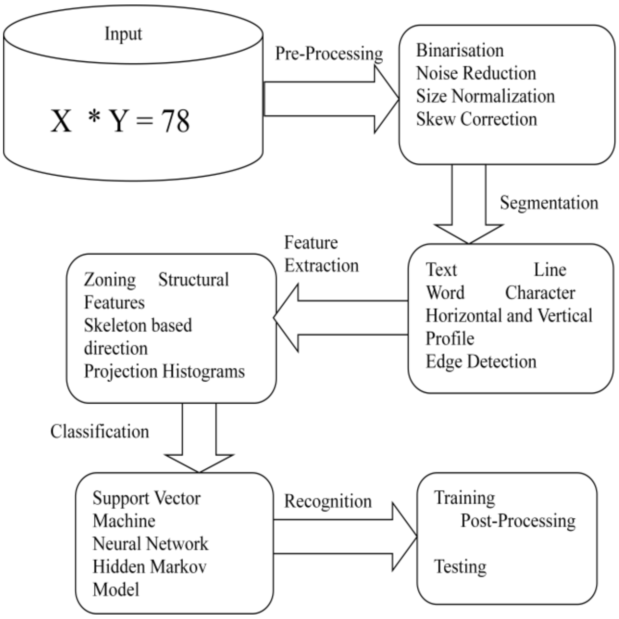


Fig 1: Architecture of Simple Mathematical Equation Recognition

Figure 1 indicates, in simple mathematical equation recognition process first image is input to system i.e. read image it may be expression or directly apply separate symbol. In next step is explained methods used for pre-processing which is apply on image to clean the image and reduce unwanted part from main object. Then we apply segmentation for separating each character. It is useful to apply feature extraction method on individual symbols. Finally we use classification technique useful to make most accurate decision for obtained feature vector. In recognition step training and testing are apply on sample to separate it into error sample and accurate samples.

This paper mainly divided into 4 sections. Section I describes introduction of recognition of simple mathematical equation. Here, each block indicates various methods useful in recognition process. Section II explains previous work related to Mathematical expression recognition, handwritten symbol recognition, on-line as well as off-line mathematical expression recognition. Section III describes steps involved in simple mathematical equation recognition system with brief explanation. In that classification and feature extraction methods are explained with recognition rate. Section IV compares various feature extraction techniques. It describes number of input samples used.

2. RELATED WORK

The history of character recognition can be traced as early as 1900, when the Russian scientist *tyurin* attempted to develop an aid for the visually handicapped. In 1929, *Gustav Tauschek* obtained a patent on OCR in Germany, followed by *Paul W.* Handel who obtained a US patent on OCR in USA in 1933. In 1935 Tauschek was also granted a US patent on his method. Tauschek's machine was a mechanical device that used templates and a photo detector. The first character recognizers appeared in the middle of the 1940s with the development of digital computers. The commercial character recognizers were available in the 1950[3]. In 1986 under thresholding processing technique *Horn* was started work on digitized images in that he proposed the gray level histogram and cumulative gray level histogram. *Dougherty* and *Giardina* was used boundary detection and edge detection technique for feature extraction in 1987. The process of shaping the image cab be acquire by 3 fundamental morphological operation that was, dilation, erosion and Skeletonization used by *Pratt* in 1991. In the period of January 2000 to July 2004 data have been collected by American mathematical society for putting handwritten or printed mathematical expression into electronic form.

3. STEPS FOR RECOGNIZING SIMPLE MATHEMATICAL EQUATION

3.1 Pre-Processing

The preprocessing is a series of operations performed on the scanned input image. It essentially enhances the image rendering it suitable for segmentation. [4] Preprocessing aims to produce data that are easy for the character recognition systems to operate accurately. [5] The main objectives of preprocessing are in the following Figure 2.

Pre-processing is the name given to a family of procedures for smoothing, enhancing, Filtering, cleaning-up and otherwise massaging a digital image so that subsequent algorithm along the road to final classification can be made simple and more accurate.

Preprocessing aims to produce data that are easy for the CR systems to operate accurately. The main objectives of preprocessing are: [6]

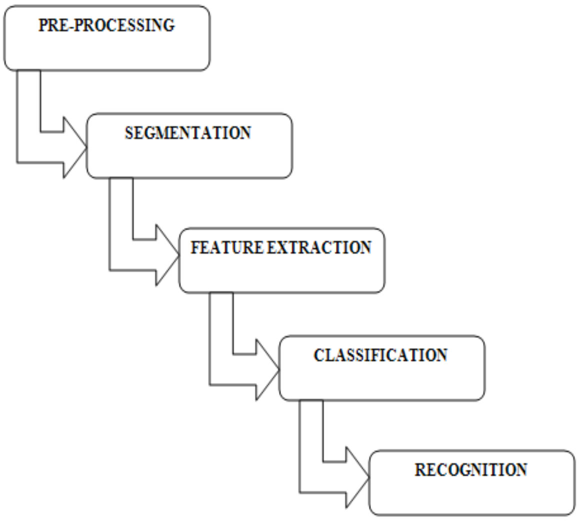


Fig 2: Phases for Mathematical Equation Recognition.

1. Binarization
2. Noise reduction
3. Size Normalization
4. Skew Correction

In order to achieve the above objectives, the following techniques are used in the pre-processing stage.

3.1.1 Filtering

This aims to remove noise and diminish spurious points, usually introduced by uneven writing surface and/or poor sampling rate of the data acquisition device. Various spatial and frequency domain filters can be designed for this purpose. The basic idea is to convolute a predefined mask with the image to assign a value to a pixel as a function of the gray values of its neighboring pixels. Filters can be designed for smoothing, sharpening, thresholding, removing slightly textured or colored background, and contrast adjustment purposes.

3.1.2 Morphological Operations

The basic idea behind the morphological operations is to filter the document image replacing the convolution operation by the logical operations. Various morphological operations can be designed to connect the broken strokes, decompose the connected strokes, smooth the contours, prune the wild points, 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 erratic hand movement [6].

3.1.3 Thinning (Skeletonization)

Skeletonization is also called thinning. Skeletonization 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 no doubt, this process reduces the processing time too. While it provides a tremendous reduction in data size, thinning extracts the shape information of the characters [6].

3.2 Segmentation

In the segmentation stage, an image of sequence of characters is decomposed into sub-images of individual character. The preprocessed input image is segmented into isolated characters by assigning a number to each character using a labeling process. This labeling provides information about number of characters in the image [4].

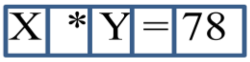
Segmentation refers to a process of partitioning an image into groups of pixels which are homogeneous with respect to some criterion. Character segmentation is a key requirement that determines the utility of conventional Character Recognition systems. It includes line, word and character segmentation. [5].

Simple equation:

$$X \quad * Y = 78$$

Fig 3: Simple Equation before Segmentation

After segmentation:



.

Fig 4: Simple Equation after Segmentation

The final result is:

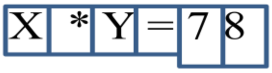


Fig 5: Final Result of Segmentation

3.3 Feature Extraction

During or after the segmentation procedure the feature set, which is used in the training and recognition stage, is extracted. Feature sets play one of the most important roles in a recognition system. A good feature set should represent characteristic of a class that helps distinguish it from other classes [3].

This objective can be achieved by following methods:

1. Zoning: in this method we can extract feature using N*M zones for essential characteristics of symbols
2. Skeleton based direction
3. Projection histogram
4. Profiles: store boundary values from four directions (top, bottom, left and right) of symbols.
5. Structural features: here crossing points, end points and loops also consider of symbols while extract the feature.

The Preprocessed Image is given as input to feature extraction module. The mean and standard Deviation will be extracted from the images. These two are called the statistical features of the histogram [7].

3.4 Classification Techniques

3.4.1 Hidden Markov Models

Hidden markov model is finite set of states, each of associated with (generally multidimensional) probability distribution. Transitions among the states are governed by set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated according to associate probability distribution. It is only the outcome not ‘the state visible to external observer and therefore states are “Hidden “to the outside, hence name is HMM. Hidden Markov models (HMM) have been widely used for MS classification in online ME recognition. However their use in offline ME recognition remains unexplored. In recent years, HMM has been successfully used for offline handwritten text recognition. In this work, it explore the technique described in applied to printed MS recognition [1]. The crucial part of scene analysis is feature extraction. A proper way to describe the given object in a knowledgeable and compact manner is the goal of this stage [8].

3.4.2 SV Classifier

Support Vector machine is one of the supervised learning method. First practical implementation of SVM had been executed in early nineties. It is most efficient family of algorithms in Machine Learning and computationally efficient. Support Vector Machines (SVM) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. This learning strategy is introduced by Vapnik and co-workers. Support vector machine is one of the best techniques used for linear and nonlinear classification and regression [3]. Therefore, it is used in recognition of handwritten en English character. The SVM classifier was originally developed for two-class or binary classification and the demanding applications of pattern recognition led to the design of multi-class SVM classifiers using the binary SVM classifiers [9].

Many classifying methods can be investigated. A template matching method is used by some systems, however this method can be slow and time consuming. Structural recognition methods are less used in mathematical expressions recognition. Systems as those in extract structural primitives and use them by comparing with the training data. On the other hand, artificial neural networks (ANN) are known to be better in terms of speed and recognition rate. Some methods perform a simultaneous segmentation and recognition such as hidden markov model (HMM) they are based on statistical models. Each symbol has its own model where recognition results are obtained as probabilities of different models [10].

3.4.3 Convolution Neural Network

In 1995, in the problems of the multilayer perceptron, was try to solve by *Yann LeCun* and *Yoshua Bengio* introduced the concept of convolutional neural networks, which were neurobiologically motivated by the outcomes of locally sensitive and orientation-selective nerve cells in the visual cortex of the cat. They designed a network structure that implicitly extracts relevant features, by restricting the neural weights of one layer to a local receptive field in the previous layer. Thus, a feature map is important factor is mainly obtained in the second layer. By decreasing the spatial resolution of the feature map, a certain degree of shift and distortion invariance is achieved. Also, the number of free parameters is significantly reduced by using the same set of weights for all features in the feature map.

To simplify process of recognition neural network is used. It is relatively simple to use and we can achieve faster result because this method avoids the segmentation before recognition. It is only possible in case, when we use quasi-holistic method using a classifier that is able to recognize two partially overlapped digits. In this work *Dan Cireson* and *Dan Pescaru* recognize numeral with overlapped digits is to address the simplest form of overlapping: the case of two digits. For this kind of classifier the input samples like 00 to 99 digits. The network architecture is combination of one hidden layer and convolution structure [11].

Table 1 depicts various approaches and methods used for mathematical symbol and expression recognition. Based on Various papers, table 1 represents study of classification, and feature extraction and segmentation techniques.

Table 1: Categories of Symbol Recognition Methods with Recognition Rate

Author	Method	No of Symbols	Recg. Rate
Francisco Álvaro, Joan Andreu Sanchez[1]	k-Nearest-Neighbor Euclidean distance Support vector machine (SVM) Weighted Nearest Neighbor (WNN) Hidden Markov models Gaussian distributions	2233 symbols: InftyCDB-1 database 25% for test and 75%for training	98.5 % (Avg)
Ahmad Montaser Awal, Harold Mouchère, Christian Viard-Gaudin[10]	multi-layer perceptions neural network (MLP)	839 symbols: including digits, Roman letters, Greek letters, binary operators.	87.5 %
Stephen M. Watt and Xiao fang Xie [12]	Without using any feature	227 symbols: including digits, Latin letters, some Greek letters and mathematical operators	94.8 %
Sajjad S. Ahranjany, Farbod Razzazi , Mohammad H. Ghassemian[13]	Convolution neural Network (Experiments with one hidden layer Perceptions)	two-digit strings, there are one hundred numbers, from 00 to 99	94.6 %

Above table describe different methods for off-line and on-line mathematical symbol recognition. Among different adopted methods, multilayer perceptron and convolution neural network produces 87.5% and 94.6% recognition rate, respectively. Also, support vector machine technique produces 98.5% accuracy which is higher than any other techniques which is depicted in the given table. Again, database utilized for support vector technique is larger than the databases used for other technique. Eventhough, it produces better results. That means, support vector machine technique is most suitable classifier for recognition of simple mathematical equations.

4. EVALUATION OF EXISTING FEATURE EXRACTION TECHNIQUES

Feature extraction is use characteristics like width, height, angle between end points, and width to height ratio which are extracted from the given samples. Every group of strokes is having different weighed which is varying from another group of samples [14].

Table 2: Comparison of Various Feature Extraction Method

Author	Method used	Remarks
Sajjad S. Ahranjany, Farbod Razzazi, Mohammad H. Ghassemian[13]	Structural analysis	Sample character 26617 50 texting experience 729 math symbol Error rate 3.84%(Max)
Widad Jakjoud, Azzeddine Lazrek[15]	Contours approach	_
Xue-dong Tian, Li-na Zuo, Fang Yang, Ming-hu Ha [16]	Gabor feature	100 Chinese mathematical literatures
Yu-sheng Guo, Lei Huang, Chang-ping Liu, Xin Jiang[17]	Structural analysis 1.Matrix analysis 2.Sub-Expression 3.Script analysis	3268 images of mathematical expression 97.2% math expression 78.2%perfect analyzed
Yu SHI and Frank K. SOONG[18]	Segmentation Hypothesis Extraction(tree search using A* search)	2579 written expression,59166 strokes and 43300 symbols
Hsi-Jim Lee and Jiumn-Shine Wang[19]	Embedded math expression extraction 4*4 non uniform blocks 4-dimentional feature vector	127-letters 36-math operator 20-numerals 7-seperator Testing symbols-3540 Error symbols-144

Table 2 illustrates various methods used for feature extraction, dataset used for recognition, testing and training samples used. From the observation of given table, it is found that structural approach is one of the trusted methods for feature extraction because it produces higher accuracy, minimum error rate and applied on large samples of mathematical symbols or equations.

5. CONCLUSION

Recognition of simple mathematical equation incorporates common steps in image processing. But due to complexity in symbols of equations, improvisation in recognition rate becomes more challenging. For that purpose, this paper evaluates the various techniques which may improve the result of recognition. Using various research papers we compared classification and feature extraction methods. Hence it is observed that support vector machine should be used as classifier and structural analysis method for extracting features from samples.

6. REFERENCES

[1] Francisco Álvaro, Joan Andreu Sanchez,” Comparing Several Techniques for Offline Recognition of Printed Mathematical symbols”, 2010 International Conference on Pattern Recognition.

[2] Qi Xiangweri Pan Yusup Wang Yang,” The study of structure analysis strategy in handwritten recognition of general mathematical expression”, 978-0-7695-3600-2/09 2009 IEEE.

[3] Nafiz Arica, “An Off-line character recognition system for free style Handwriting” thesis submitted on Sep 1998.

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

[5] http://www.scribd.com/doc/60245721/English-Character-Recognition-System-Using-matlab.

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

[7] Shubhangi D.C., P.S. Hiremath, “Multi-Class SVM Classifier for English Handwritten Digit Recognition using Manual Class Segmentation” International Conference on Advances in Computing, Communication and Control, 2009 ACM

[8] Mou-Yen Chen, Amlan Kundu and Sargur N. Srihari, Fellow, “Variable Duration Hidden Markov Model and Morphological Segmentation for Handwritten Word Recognition”, IEEE transactions on image processing, vol. 4, no. 12, December 1995.

[9] Shailedra Kumar Shrivastava, Sanjay S. Gharde, “Support Vector Machine for Handwritten Devanagari Numeral Recognition” International Journal of Computer Applications (0975 – 8887), Volume 7– No.11, October 2010.

[10] Ahmad-Montaser Awal, Harold Mouchère, Christian Viard-Gaudin,” Towards Handwritten Mathematical Expression Recognition” 2009 10th International Conference on Document Analysis and Recognition.

[11] Dan cirasan, Dan pescaru,” Off-line Recognition of Handwritten Numeral Strings Composed from Two-digits Partially Overlapped Using Convolutional Neural Networks”, 978-1-4244-2673- 7/08 2008 IEEE.

[12] Stephen M. Watt and Xiao fang Xie, “Recognition for Large Sets of Handwritten Mathematical Symbols”, 1520-5263/05 2005 IEEE.

[13] Sajjad S. Ahranjany, Farbod Razzazi, Mohammad H. Ghassemian,” A Very High Accuracy Handwritten Character Recognition System for Farsi/Arabic Digits Using Convolutional Neural Networks”, 978-1-4244-6439-5/10 2010 IEEE.

[14] George Labahn, Edward Lank, Scott MacLean, Mirette Marzouk, David Tausky,” MathBrush: A System for doing Math on Pen-Based Devices”, 978-0-7695-3337-7/08 2008.

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

[16] Xue-dong Tian, Li-na Zuo, Fang Yang, Ming-hu Ha,” An Improved Method Based On Gabor Feature for Mathematical Symbol Recognition “, -4244-0973-X/07 2007 IEEE.

[17] Yu-sheng Guo, Lei Huang, Chang-ping Liu, Xin Jiang” An Automatic Mathematical Expression understanding System”, Institute of automation, Beijing 100080, china.

[18] Yu SHI and Frank K. SOONG “A symbol Graph Based Handwritten Math Expression Recognition “, 978-1-4244-2175-6/08 2008 IEEE.

[19] Hsi-Jim Lee and Jiumn-Shine Wang, “Design of a Mathematical Expression Recognition System”, 0-8186-7128-9/95 1995 IEEE.