International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-6, May 2013

Design And Implementation of Special Symbol Recognition System using Support Vector Machine

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

Abstract: Recognition of printed mathematical symbols is a subject of growing interest to automatically convert scientific paper documents into electronic form. Several methods have been proposed for recognition of printed and handwritten symbols. Symbol recognition in mathematical expressions is also difficult because there is a large character set with a variety of font styles and font sizes. Generally, in many applications, it is necessary to copy the contents from some original documents which may be in PDF like format. While accessing the data from that document, if it encounters symbols it remains unread in the copied document. So it is very difficult to read the original document. In proposed work, Discrete Cosine Transform method is used to extract the features of the symbol and Support Vector Machine, which performs the classification task used as classifier. Support Vector Machine provides the high accuracy at the time of classification.

Index Terms: Symbol Recognition, Discrete Cosine Transform, Support Vector Machine.

I. INTRODUCTION

A symbol represents an idea, a process or a physical entity. The main purpose of a symbol is to communicate meaning [1]. Numerals are symbols for numbers. Personal names are also symbols representing individuals, written languages are composed of a variety of different symbols that create words. Symbol represents the meaningful information to user. Each and every symbol is having different meaning. Some Symbols are used to form mathematical equation. Mathematical symbols are of more useful in mathematical equation, scientific documents and to create engineering drawings.

Various techniques have been proposed, during the last 10 years, for recognizing symbols in documents. Differences depend on both application fields. Detecting and classifying symbols is one of the most relevant aspects of analyzing many types of documents [2]. Different types of symbols appear in most of the scientific & engineering document. Following are some areas where the need of symbol recognition could arise:

- Mathematical inputting and editing have faced problems due to their two dimensional structure and large sets of symbols. These typically consist of special symbols and Greek letters in addition to English letters and digits. The commonly used keyboard input is thus insufficient for the input of such a large set of symbols, which has led to the desire for other input methods [3].
- While copying the contents from any PDF like document, instead of some symbol, garbage value would copied to that location. So it should be necessary to identify such symbols.
- Recognition of mathematical symbols could provide a user friendly interface for computer algebra systems.

The main aim is to recognize all the symbols correctly and accurately, so that society people can use this system. Symbol recognition is one kind of character recognition. Character recognition is a process, which assigns a symbolic meaning with the objects like letters, symbols and numbers that are on an image[4]. Each symbol having different meaning so it is obvious to use different technique to handle or recognize different symbols. Symbol recognition is an important area in computer vision that involves the identification of symbols in an image or video. Symbol recognition is used in a large number of different applications such as [5] interpreting and converting scanned engineering drawings and circuit diagrams into other electronic document formats, querying images from databases based on shape and recognizing characters and words within an electronic document.

There are some issues that make symbol recognition a difficult problem to solve. Symbols typically contain little or no color or texture information that can be used to distinguish between different symbols. Under such circumstances, a shape based similarity method is needed to perform symbol recognition[5].

II. RELATED WORK

Table I. gives the overview of various Symbol recognition systems. Different author presented paper on different types of symbols like handwritten symbol, printed symbol, Graphic symbol, drawing symbol and sketched symbol. Each of them used different classification techniques like Hidden Markov model, Neural Network, Decision tree classifier, Nearest

Manuscript published on 30 May 2013.
***Correspondence Author(s)**

Sanjay S. Gharde, Compter Engineering Dept., SSBT’s COET, Jalgaon, India.
Vidya A. Nemade, Compter Engineering Dept., SSBT’s COET, Jalgaon, India.
K.P.Adhiya, Compter Engineering Dept., SSBT’s COET, Jalgaon, India.

TABLE I. OVERVIEW OF SYMBOL RECOGNITION TECHNIQUE

Sr.	Author	Classification Techniques	Recg. Rate
1	GONG Xin, PE Jihon ,	Hidden Markov model	85%
2	T. Cheng, J. Khan, H. Liu & D. Y. Y. Yun [7]	Neural Network	89%-Testing 98.5%-Training
3	Jisheng Liang, Vikram Chalana, Robert Haralick [8]	Decision tree classifier, Nearest Neighbor	Moment Invariant-82.3% Zernike Moment-86.9% Probability Map-95.7%
4	Xue-Dong Tian, Li-Na Zuo, Fang Yang, Ming-Hu Ha [9]	Minimum Distance Classifier	High Quality-97.8% Low Quality -97.1%
5	Heloise Hse, A. Richard Newton[10]	Support Vector Machine	96.7%
		Neural Network	94.1%

Neighbor Minimum Distance Classifier and Support Vector Machine. Xue-Dong Tian [9] achieved recognition rate of 97.81%.

The goal is to detect ,extract and segment the symbol. So in proposed work, Support Vector Machine (SVM) is used as classifier, which gives the high accuracy at the time of classification and Discrete cosine Transform (DCT) method is used to extract the feature of symbol.

This paper proposes a methodology for the special symbol recognition form English text. The preprocessed image is segmented into individual characters. The features of the symbols are extracted using Discrete Cosine Transform (DCT). The DCT transforms a signal from the spatial domain to the frequency space, with minimum information redundancy. These features are feed to Support Vector Machine for performing classification and regression task.

The rest of the paper is organized as follows. In section III, the proposed symbol recognition system is presented. Section IV presents the experimental results and discussion. And finally, the paper is concluded in section V.

III. THE PROPOSED SYMBOL RECOGNITION SYSTEM

In this section, the proposed recognition system is described. The typical character recognition system consists of preprocessing, segmentation, feature extraction and classification. The general schematic diagram of the proposed symbol recognition system is shown in Fig.1.

A. Preprocessing

The pre-processing is a series of operations or steps performed on the scanned input image. It basically enhances the image rendering it suitable for segmentation. The various tasks performed on the image in preprocessing stage are Binarization, noise removal, morphological operation. e.g. Binarization process converts a gray scale image into a binary image using global thresholding technique. Some of the preprocessing operation are described as follows:

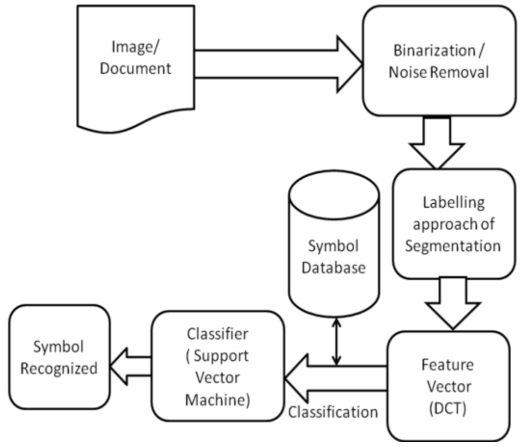
Image/ Document
Binarization / Noise Removal
Symbol Database
Labelling approach of Segmentation
Symbol Recognized
Classifier (Support Vector Machine)
Classification
Feature Vector (DCT)

Fig.1. Schematic diagram of the proposed symbol recognition system.

1) Noise Reduction

The noise which is introduced by the optical scanning devices or the writing instrument causes disconnected line segments, bumps and gaps in lines, filled loops etc. The distortion which includes local variations, rounding of corners, dilation and erosion, is also a problem. Prior to the character recognition, it is necessary to eliminate these imperfections. There are many techniques to reduce the noise present in image like filtering and morphological operations.

The basic idea behind the morphological operation is to filter the document image. Various morphological operations 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 effectively used to remove the noise on the document images due to low superiority of paper and ink, as well as irregular hand movement [11].

2) Compression

Two popular compression techniques are thresholding and thinning. Thinning extracts the shape information of characters or symbol. In character recognition most of the information can be extracted from the shape of the strokes. Therefore, the skeleton of a given character is necessary and in most there are two basic approaches for thinning. Pixel wise thinning and Non-pixel wise thinning.

In order to reduce storage requirements and to increase the processing speed, it is often desirable to represent grey scale or color images as binary images by selecting some threshold value for everything above that value is set to 1 and everything below is set to 0. The process of converting a grey-level image to a binary image is called thresholding or binarization [12]. Currently there are two broad categories of thresholding algorithms: 1) Global and 2) Locally adaptive methods.

B. Segmentation

It is an operation that decomposes an image of sequence of characters into sub images of individual symbols. After accepting the document, the document image is subjected to preprocessing for background noise elimination and skew correction to generate the bit map image of the text. The preprocessed image is then segmented into lines, words and characters or symbols [13].

The purpose of the segmentation is to extract each character from the text present in the image [14]. The process of segmentation for the proposed work mainly follows the following pattern:

- First, it identifies the page layout, identifies the line from the page, identifies the word from that line, and finally, identifies the character from that word.
- Starts from the first pixel still it finds the continuous black pixel from left to right direction.
- After that if white pixel found then this indicates the one character or symbol from the image.
- Apply bounding box on that symbol or character image.

C. Feature Extraction

After preprocessing on the image of text, features of character or symbols are extracted. This step is heart of the system. This step helps to classify the characters based on their features. Feature extraction is the name given to a group of procedures for measuring the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure [15]. The feature extraction stage analyses a text segment and selects a set of features that can be used to exclusively identify the text segment. The issue of choosing the features to be extracted should be guided by the following concerns:

- The features should carry sufficient information about the image and should not necessitate any domain-specific knowledge for their extraction.
- They should be easy to calculate in order for the approach to be feasible for a large image collection and rapid retrieval.
- They should be related well with the human perceptual characteristics because users will finally decide the correctness of the retrieved images [16].

For the proposed system, Discrete Cosine Transform method is used to extract the features. The Discrete Cosine Transform converts a continuous signal (data of the image) into its elementary frequency components. Because the DCT of an image captures the most visually important information in just a few coefficients, it is widely used in image algorithms such as compression. The DCT is a popular signal transformation method, which is makes use of cosine functions of different frequencies, as kernels.

The following are some of the properties of the DCT

1. The DCT is real.
2. The DCT is a fast transform. That is, the time complexity of the DCT is comparable to that of the Fast Fourier Transform (FFT).
3. The DCT concentrates the majority of the image energy in very few coefficients.

It stores the lowest frequency components in upper left corner, whereas the highest frequency components in bottom right corner [17], as shown in Fig.3. The capability of the DCT is to compress energy makes the DCT suitable for pattern recognition applications. The DCT II, is normally used in image processing and compression (JPEG, MPEG), because it has strong energy compaction, meaning that a few coefficients enclose the most of the signal in process.

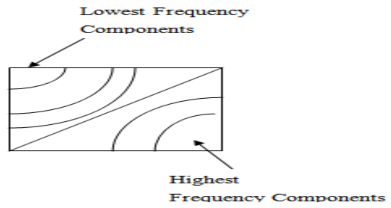

Fig. 3. 2-DCT frequency distribution [17]

D. Classification

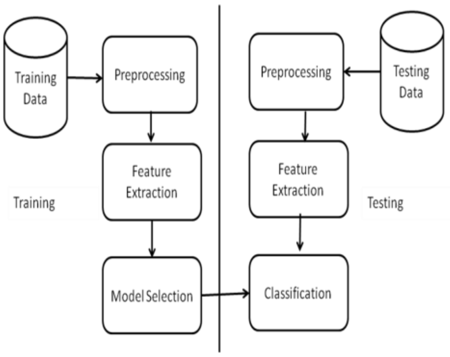

Fig. 4. The General Classification Process [19]

Classification is used as the decision making stage of recognition system. It uses the features extracted in the previous stage to identify the text segment according to preset rules [18]. Many types of classifiers are applicable to OCR like K-Nearest Neighbour, Hidden Markov Model, SVM etc. For the proposed work SVM is used as the classifier. There are two steps in building any type of classifier: training and testing as shown in Fig.4. These steps can be broken down further into sub-steps [19]:

1) Training

Preprocessing – Processes the data so it is in a appropriate form for further processing.

Feature extraction – Reduce the amount of data by extracting significant information. Generally results in a vector of scalar values.

Model Estimation – From the finite set of feature vectors there is need to estimate a model for each class of the training data.

2) Testing

Preprocessing- Process the data.

Feature extraction – Extract the features of the data.

Classification – Compare feature vectors to the different models and find the closest match [19].

3) Support vector Machine

The SVM (Support Vector Machine) was introduced first by Vapnik and co-workers in 1992 [20]. Support Vector Machines are a group of supervised learning methods which can be applied to classification or regression. The SVM classifier accepts the set of input data and predicts to classify them in one of the only two distinct classes. Basically, SVM classifier is trained by a given set of training data and a model is prepared to classify test data. Depending on how all the samples can be classified in dissimilar classes with appropriate margin, different types of kernel in SVM classifier are used. Frequently used kernels are: Linear kernel, Polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid (hyperbolic tangent) [21].

*Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)*

IV. RESULTS AND DISCUSSION

For the proposed work, 84 symbols are considered for the experiment. MATLAB 7.0 is used for implementation. The prototype developed is extensively tested on different images. We considered 400 samples of the data and tested on different text images of Times New Roman font. The Support Vector Machine classifier is optimizing an error function that minimizes the misclassification on the training set.

TABLE II. RECOGNITION RESULT

Sr. No.	Input	No. of symbols in the image	No. of Recognized symbols	Recognition Rate %
1	Test1	54	54	100
2	Test2	44	44	100
3	Test3	28	28	100
4	Test4	29	28	96.55
5	Test5	39	38	97.43
6	Test6	30	30	100
7	Test7	35	34	97.14
8	Test8	50	50	100
9	Test9	52	51	98.07
10	Test10	20	20	100
Average Recognition Rate				98.91%

Table II. Shows the Recognition results of identifying symbols from a document image with a different font size. Such that Test1,Test2, Test3 contains the text with font size of 16, Test4, Test5, Test7 contains the text with font size 12. Like that rest of the others contain image with font size of 14. For all such images proposed system provides the recognition rate of 98.91% as shown in Table2.

V. CONCLUSION

The purpose of this work is to identify symbols from a document image containing both texts and symbols. The presented symbol recognition system can identify symbols from a document image through examining some distinct feature of the preprocessed images of English Text. DCT feature values of symbol-images have been taken after a series of preprocessing and segmentation operation. Total 400 samples of the different 81 symbols are considered for the experiment. Support Vector Machine which performs the classification and regression provides 98.91% of recognition accuracy. Further it can be extended for the recognition of word, scientific mathematical expressions since we primarily recognize special symbols only.

VI. ACKNOWLEDGMENT

The authors honourably thankful to Computer Engineering Department of SSBT's, College of Engineering & Technology, North Maharashtra University (NMU) for the support in making this projects success.

REFERENCES

1. http://en.wikipedia.org/wiki/Symbol
2. L. P. Cordella, M. Vento, "Symbol recognition in documents: a collection of techniques?" International Journal on Document Analysis and Recognition (IJDAR),pp 73-78, Springer Varlag 2000.
3. Xiaofang Xie , "On the Recognition of Handwritten Mathematical Symbols", Thesis submitted at London, Ontario, Dec 2007.
4. http://www.scribd.com/doc/60245721/English-Character- Recognition -System-Using-matlab.
5. Alexander Wong and William Bishop, "Robust Hough-Based Symbol Recognition Using Knowledge-Based Hierarchical Neural Networks".
6. GONG Xin, LI Cuiyun, PE Jihon, XIE Weixin, "HMM Based Online Hand-Drawn Graphic Symbol Recognition",ICSP'02 Proceedings,pp-1067-1070, 2002 IEEE.
7. T. Cheng, J. Khan, H. Liu and D. Y. Y. Yun, "A Symbol Recognition System",pp-918-921, 1993 IEEE.
8. Jisheng Liang' Ihsin T. Phillips Vikram Chalana' Robert Haralick, "A Methodology for Special Symbol Recognitions", 2000 IEEE.
9. Xue-Dong Tian, Li-Na Zuo, Fang Yang, Ming-Hu Ha," 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, IEEE.
10. Heloise Hse, A. Richard Newton, "Sketched Symbol Recognition using Zernike Moments", Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) IEEE.
11. Nafiz Arica and Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on of Character Recognition Focused on Off-Line Handwriting" pp 216-233, IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, VOL. 31, NO. 2, MAY 2001 IEEE.
12. Myer Blumenstein, "Intelligent Techniques for Handwriting Recognition" Thesis submitted in December 2000.
13. Nasien ,Habibollah Haron, Siti Yuhaniz, "Support Vector Machine for English handwritten character recognition" 2010 second International conference on Engineering and applications 2010 IEEE.
14. Anita Pal & Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network",, International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 141-144.
15. G.A. Papakostas, D.E. Koulouriotis and E.G. Karakasis, "Efficient 2-D DCT Computation from an Image Representation Point of View" ,pp 21-34.
16. Daniela Stan Raicu, " Tutorial 2: Image Feature Extraction" Visual Computing Workshop: Image Processing , DePaul University May 21, 2004
17. Sai Charan K., "A Block DCT based Printed Character Recognition System " Dissertation submitted at Prashanthi Nilayam, March 2006.
18. J.Pradeep, E.Srinivasan, S.Himavathi, "Neural Network based Handwritten Character Recognition system without feature extraction", pp 40-44, ICCCET 2011,March. 2011 IEEE.
19. Jesse Hansen, "A Matlab Project in Optical Character Recognition (OCR)
20. Shailedra Shrivastava , Sanjay S. Gharde "Support Vector Machine for Handwritten Devanagari Numeral Recognition" International Journal of Computer Applications Oct 2010.
21. Sukhpreet Singh, Ashutosh Aggarwal, Renu Dhir, "Use of Gabor Filters for Recognition of Handwritten Gurmukhi Character", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5, May 2012, pp 324-240.