Diagnosis of Mathematical Symbols using Hidden Markov Model

Mohamad Hassan Asadi
Islamic Azad university of Larestan, Larestan,Iran

Abbas Akkasi
Islamic Azad university of Larestan, Larestan,Iran

Ebrahim Zargarpour
Islamic Azad university of Larestan, Larestan,Iran

Zahra Mohammdi
Islamic Azad university of Larestan, Larestan,Iran

ABSTRACT

Diagnosis of mathematical symbols in handwritings is originated from Optical Character Recognition (OCR) method. Recognition of mathematical symbols increases the accuracy of calculations. In present study, hidden Markov model is applied with a new feature selection system. Considering previous studies, a lot of researches performed on mathematical symbols recognition, have used support vector machine. Test process in this method is time-consuming and it is not advised to use it. In this new approach, the result is 96.05% accuracy for Infity database and 96% for IRISA database.

Keywords

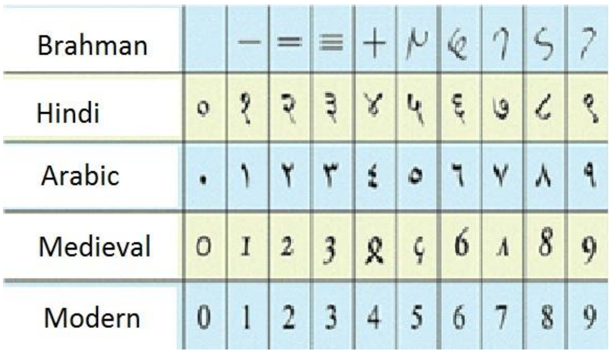
Mathematical symbols, optical character recognition (OCR), hidden Markov model

1. INTRODUCTION

Mathematics is defined as a science of study on structural pattern, transition and space; in an unofficial definition, it is described as the science of facts and figures. Its definition has changed based on the extent of its range and also expanded scope of mathematical thinking.

Mathematics has its own language form where words and symbols in writing are replaced by facts and figures. In thinkers' points of view, verification of evident abstract structures is applied by logic and mathematic symbolization.

Another important point in mathematics is that numbers are used in addition to symbols. In a way that mathematics has no meaning without them. The method of number writings is diverse in different languages. Table 1 shows some examples of number writing form in different languages. Recognition of these forms is not mentioned in present study.



Brahman		—	=	≡	+	𑁖	𑁗	𑁘	𑁙	𑁚
Hindi	०	१	२	३	४	५	६	७	८	९
Arabic	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Medieval	0	I	2	3	ℛ	ς	6	ʌ	8	9
Modern	0	1	2	3	4	5	6	7	8	9

Table 1: some examples of number writing forms in different languages

Recognition of numbers in different languages has been studied a number of times e.g. in Persian language including Soltanzadeh and Rahmati (2004), Azmi and Kabir (2001), Dehghan (2001) and Nabavi et al. (2005). Nearly, all of these studies are not based on standard gathered data.

Al-Omari et al. [9] used Possible Neural Network (PNN) for Arabic numbers. Their database contained 720 data points with 99.75% accuracy. Drucker et al. [11] worked on a text based on SVM with 97.03% accuracy.

The preliminary set of numbers were engraved on sticks and called "tally". Special symbols were allocated to these sets (2, 5 and etc.) and a set of calculus was created. Mathematicians ordained distinctive symbols including summation and equality symbols and also invented special words in order to define new conceptual. Some of these symbols are shown in Table 2.

Table 2. Some of mathematical symbols

Name	Symbol	Name	Symbol
Summation	+	Subtraction	-
Radical (Square Root)	√	Parentheses (Grouping priority)	( )
equality	=	Multiplication	×
Plus-minus	±	Smaller/Greater than	> <
Radical (nth Root)	$\sqrt[n]{x}$		
Power	x^y		

In present paper, the advantage of a new approach will be taken in feature extraction using Hidden Markov Method (HMM). HMM is a common method in pattern recognition in different studies. In the second stage, HMM will be discussed and mentioned structure will be discussed in the third section and finally an identification model will be proposed and the results will be presented.

2. HIDDEN MARKOV MODEL

The HMM method is based on adaption of image formats to a string of a doubly stochastic model. They have a clear and immediate application in signal processing and its recognition and respected signal is naturally presented as a sequence of spectral estimates varying with time.

HMM classifier is composed of a set of definite possible states Q ={q1,Λ, qk } and possibility of transition between these states. Suppose we have k possible states and each state has a dispersion probability distribution in a form of p(x| qk).

The possibility of p(x| qk) is usually considered using the combination of Gaussian probability density function as the possibility of HMM dispersion [4]. One method for training Gaussian parameters is to maximize the similarities; it means that combinations will be trained for each class data points, not for separation of one class from the others.

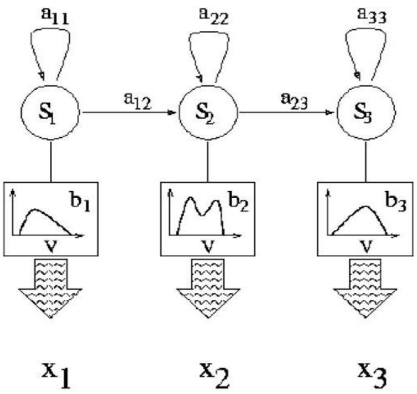


Figure 1: HMM with Gaussian combination [1]

3. PROPOSED METHOD OF RECOGNITION

HMM method is used alone for recognition and a new model for segmentation will be suggested in a more optimized way. First of all, a pre-process is applied on the symbols to remove noises and make a better image for recognition.

The outline of this method has originated from [1] and the model is like Figure 2. Completing the pre-process stage and classification of data points, a number of assumptions are obtained. These assumptions will be identified through Gaussian Markov model.

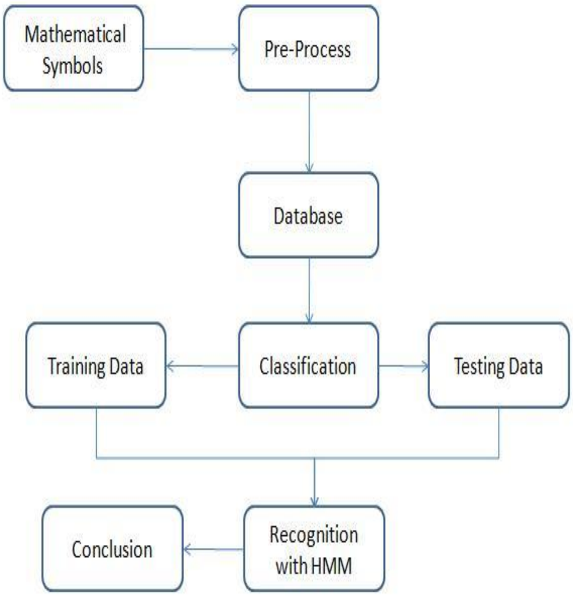


Figure 2: suggested model for recognition of mathematical symbols

During classification, each image will be divided into 9 sectors. Applying this method reduces the number of comparison states dramatically and prevents overload of Markov model.

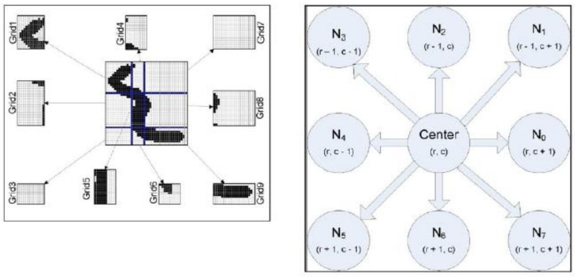


Figure 3: image classification method [1]

While hidden Markov model is used, calculated possibilities are defined as input in this model and hidden model performs recognition procedure by these possibilities.

In present model, 'S' stands for states and 'anm' is the possibility of dispersion transition between different states.

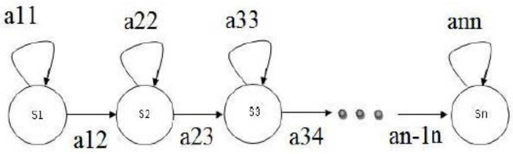


Figure 4: Hidden Markov Model

4. CONCLUSION AND DATABASES

In present study, two databases (Infity and IRISA) used for identification of mathematical symbols. Infity database was available in two versions and this study carried out on the latest version of InfityCDB-2.

InfityCDB-2 is a database with various symbols in English, French and German. The numbers of these symbols are 662142, 37439 and 77812 in English, French and German, respectively. English symbol recognition was investigated in this paper. It should be mentioned that difference in symbols of languages are usually related to variance numbers and abbreviations.

The first draft of this database included 476 pages with all mathematical symbols only in English. In the second version, formulas and characters added up to the set. A newer version (e.g. InfityCDB-3) is more focused on separation of characters and symbols and no other new thing was added to the set. Figure 5 shows an example of symbols in this database.

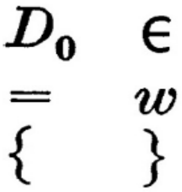


Figure 5: an example of Infity database

IRISA database contained 33150 samples. Separation between different languages is not considered in this database, but handwritten signs are included in the database as well as printed signs. Figure 6 shows an example of this database.

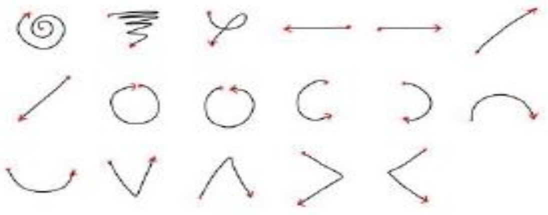


Figure 6: an example of IRISA database

Both databases have a demo version on the website and interested researchers may use it for further references.

Investigating on different method for both databases revealed the following results:

According to Table 3 for Infity database, all the results for identification using different methods were suitable, but noises can oscillate the results and reduce their efficiencies. In this database, with proposed method, for usual and noiseless numbers 96.05% accuracy was observed and for noisy states or damaged images 94.9% accuracy was achieved. In addition, this method performs faster and less complex in comparison with other methods.

Table 3: recognition results for Infity database

Method	Recognition Percentage with 20% Noise	Recognition Percentage without Noise
Hidden Markov Model (HMM)	94.9%	96.05%
SVM [13]	60%	78%
SVM [14]	59%	60%
Hierarchical Classifier [15]	94.1%	95.6%

In IRISA database (Table 4) appropriate results were attained according to different classifications. The results of this database are improved and accuracy of 96% for noiseless data points achieved, while our suggested method performs better with less training time.

Table 3: recognition results for IRISA database

Method	Recognition Percentage with Noise	Recognition Percentage without Noise
Hidden Markov Model (HMM)	94.9%	96.05%
Fuzzy [16]	86%	95%

5. REFERENCES

[1] Sameh M. Awaidah and other. A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models.signal processing, In press, 2009.

[2] C.-L.Liu and C.Y.Suen. A new benchmark on the recognition of handwritten bangla and farsi numeral characters.Pattern Recognition, In press, 2008.

[3] W.M. Pan, T.D. Bui, and C.Y. Suen : Isolated Handwritten Farsi numerals Recognition Using Sparse And Over-Complete Representations, 2009 10th International Conference on Document Analysis and Recognition

[4] Yasemin Altun and Ioannis Tsochantaridis and Thomas Hofmann: Hidden Markov Support Vector Machines, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[5] Christopher M. Bishop: Pattern Recognition and Machine Learning, 2006 Springer Science□ Business Media, LLC

[6] Usama Fayyad: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, 121–167 (1998)

[7] Ahmad A.R, Viard-Gaudin, C. Khalid M: Lexicon-based Word Recognition Using Support Vector Machine and Hidden Markov Model, 2009 10th International Conference on Document Analysis and Recognition

[8] F. Solimanpour, J. Sadri, C.Y. Suen, Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language, in: Proceedings of the 10th International Workshop on Frontiers of Handwriting Recognition, La Baule, France, 2006, pp. 3–7.

[9] Al-Omari, F., Al-Jarrah, O.: Handwritten Indian numerals recognition system using probabilistic neural networks. Adv. Eng. Inform. 18(1), 9–16 (2004)

[10] Said, F., Yacoub, R., Suen, C.: Recognition of English and Arabic numerals using a dynamic number of hidden neurons. Proc. 5th ICDAR, pp. 237–240, 1999

[11] H. Drucker, B. Shahrary, D.C. Gibbon, “Support vector machines: relevance feedback and information retrieval” ,Information Processing and Management 38, p305-323, 2002

[12] Sherif Abdleazeem and Ezzat El-Sherif: Arabic handwritten digit recognition, IJDAR (2008) 11:127–141

[13] Birendra Keshari and Stephen M. Watt, Hybrid Mathematical Symbol Recognition using Support Vector Machines, IJDAR (2007)

[14] Christopher Malon and etc, Mathematical symbol recognition with support vector machines, Pattern Recognition Letters 29 (2008)

[15] Jason Ranger and etc, Optical Character Recognition of Printed Mathematical Symbols using A Hierarchical Classifier, IPCV (2012)

[16] Abdullah Almaksour and etc, Optical Personalizable Pen-Based Interface Using Life-Long Learning, International Conference on Frontiers in Handwriting Recognition (ICFHR), Aug 2010, Kolkata, India. pp.188-193, 2010.