Recognition of Off-line Printed Arabic Text Using Hidden Markov Models

Atheel Sabih Shaker
Dept. Of Computer Science/College of Economic Sciences/ University of Baghdad
Atheel.Sabih@baghdadcollege.edu.iq

Received in:2/January/2018, Accepted in:24/January/2018

Abstract

In this paper, we introduce a method to identify the text printed in Arabic, since the recognition of the printed text is very important in the applications of information technology, the Arabic language is among a group of languages with related characters such as the language of Urdu , Kurdish language , Persian language also the old Turkish language " Ottoman ", it is difficult to identify the related letter because it is in several cases, such as the beginning of the word has a shape and center of the word has a shape and the last word also has a form, either texts in languages where the characters are not connected, then the image of the letter one in any location in the word has been Adoption of programs ready for him A long time. In this paper we present an off-line system to recognize printed Arabic text by using Hidden Markov Model with the aid of algorithm that segment the text line into sections and then into characters.

Keywords: Automatic identification, Hidden Markov series, Printed text, Automatic recognition of printed text, Markov series automatic recognition.

Introduction

Artificial intelligence is one of the most important fields of applied science in computer science. There are many applications in this field, including natural language processing, automatic translation, pattern recognition, etc.

In the recent years, the research on the field of automatic character recognition has expanded significantly in both and indirect terms. The use of the term "understanding" is not intended, as has been the case so far, Written by computer, this means trying to decode the message to be reported. [2]

"The first successful effort in this regard was due to the Russian scientist TYURIN in 1900, followed by the attempts by FOURIER DALBE to manufacture the reader machine for the 1912 speaking letters, and the prosthetic prosthesis Thomas built "THOMAS" in 1926. [3].

This paper is used in the administrative processing of administrative files, such as contracts, birth certificates, questionnaires, bank files and postal addresses. It has made great strides in foreign languages, but its applications in Arabic, despite the commendable efforts of some competent authorities, remain below the required level.

One of the applications of artificial intelligence software is to distinguish patterns [4]

Pattern recognition is also a study of how machines can observe the environment, learn to show patterns that they wish to distinguish, and make a reasonable decision about the types of patterns. [5]

As an antidote to chaos, as an undefined entity, it is possible to give a certain name. After definition [6]

Despite gradual improvements in the applications of pattern recognition in the late twentieth and early twentieth centuries, character recognition remains one of the most important issues of pattern recognition. [7]

These applications include reading the mailing address on the envelope, archiving and retrieving the text, digitizing the libraries, etc. And the distinction of patterns visually passes through several stages and the last stage is discrimination where there are several ways to conduct, and we will use in this research model in the distinction of printed Arabic text. Markov's Hidden Markov Model-HMM (Markov) is one of the models used in speech and language processing. [8] The double hidden HMM is known in which hidden cases can be viewed only by certain observations. [9]

A model for character recognition system

The character recognition system generally consists of four basic stages illustrated in Figure 1, where you begin by inserting the document containing the text we want to distinguish and ending with the characterization of the entered document. [10] As shown in Figure (1)

Figure (1): A model for character recognition

The pattern recognition system may not have all of these stages. It is possible to shorten some stages without affecting this, On the process of pattern recognition, for example, the system discriminates without requiring the stage of Features Extraction, and is used instead (Matching templates)

Markov chains

Mathematical models may be specific or coincidental. However, in many cases in life, there are coincidental phenomena (phenomena that are not completely deterministic or unpredictable in their future behavior and are termed coincidences). [11]
The cross-model becomes the most appropriate to represent it.,The system described in Figure (2) can be described over a specified period of time, as described in
(S1, S2, ..., SN) (Discrete states) (N) is one of the set of these cases

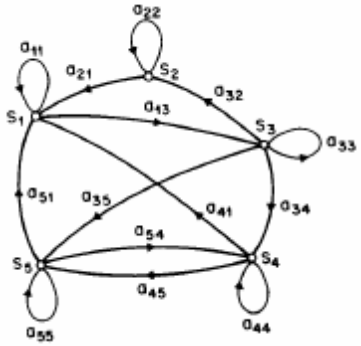

Figure (2): Markov series for (5) cases with their transitions [11].

The Hidden Markov Model

The hidden Markov model is a system of limited machine stations that depends only on its previous state at time t, which is capable of generating observations of the probability of a state transition at time. The sequence of the situation that produces the given observation is unknown. [12] Thus, in the hidden Markov model, t-1 is Time
The case is not visible, so the hidden Markov model and transitions between situations are governed by a set of probabilities called the probability of transition from a given situation that can result in a result or observation and according to the probability distribution associated with that state. [13].
The difference between the hidden Markov model and the Markov model is the existence of additional probabilities. This is the hidden part of the model and is associated with the resulting viewing of each case [14]. Markov's hidden model is a statistical model capable of statistical classification. It has therefore been applied in voice recognition and handwriting recognition because of its adaptability and versatility in the processing of chain signals [16].
As follows: [9] (HMM) The Hidden Markov model (HMM) elements define the number of cases in the model. Although the cases are hidden, many natural applications in N (S) often have some relevance to the cases or set of cases of the model the situation is as follows:
S = {S1, S2, ..., SN}
(qt) the case at time (t)
Number of single status view codes. One-view codes can be represented as follows: M•
V = {v1, v2, ..., vM}
(A) Probabilistic distribution of the transition situation
A = {aij}
Where
aij = p [qt + 1 = Sj \ qt = Si], 1 £ i, j £ N
j Probable distribution of the view code when the case
B = {bj (k)}

The proposed model using the hidden Markov model

The proposed model using the hidden Markov model as shown in, figure (3).

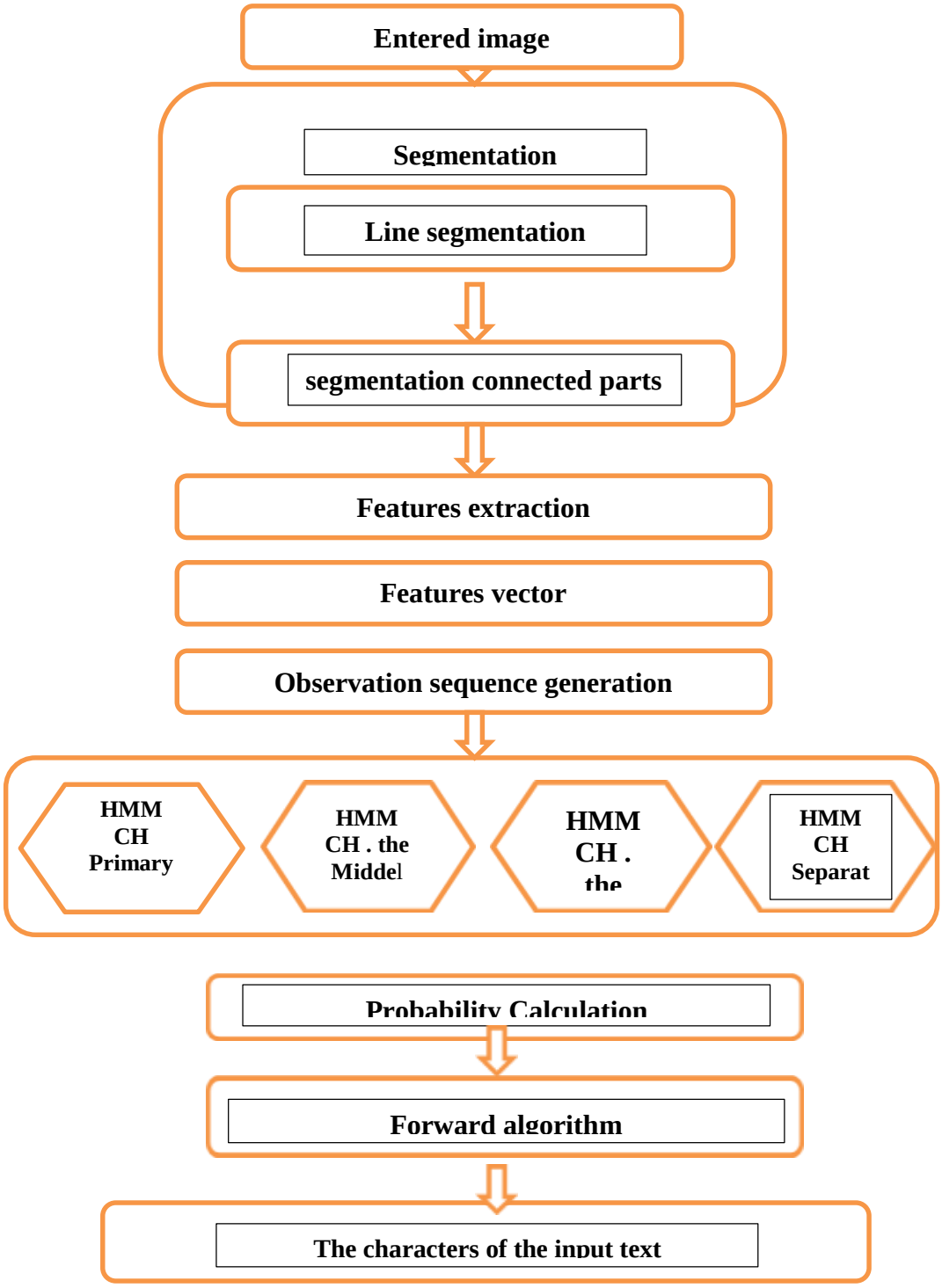

Figure (3): Proposed Scheme of recognition system

Steps to implement the proposed character recognition system-

In this research, an algorithm was proposed to distinguish the Arabic text printed using the Hidden Markov Model. Where programming was done using Matlab V 7.6 to make (Recognition)

training phase

The system begins with the training of hidden Markov models designed and included the following steps

Insert the image

This step was printed (28) lines of text per line printed, So that each line of text includes a certain character in all its forms at the beginning, medal or end, and then storing it as a binary image in the (BMP) file, the data in the image is stored in binary format (1,0), the black dot that is part of the pattern is represented by the value 0 and the white point is in value 1.
We did not perform the noise reduction because the image was not inserted by optical scanning devices such as a scanner or a penlight that causes noise

Figure (4): is a letter (ب) in its four forms

Cutting stage

The stage of cutting is an important stage within the stages of the system of distinguishing the Arabic text because of the nature of Arabic writing
Which require separating patterns of character patterns from one another. Two-step automatic shredding is performed. The text line is cut into words and / or sections using the vertical diagram. Each word and / or clip is then cut into its constituent characters
The character assignment is done after the start and end of the characterization process as well as finding:
1- Base line: Be at the line that has the largest number of black dots.
2- Top line: For each column in the section.
3- Bottom line: For each column in the section.
4- Threshold: Corresponds to the largest duplicate value in the histogram for each column created in the previous chipping step.
5- Number of vertical transitions (0, 1) and (1, 0)
The end column must meet several conditions. The start column of the character has a histogram that is larger than the threshold.

Features extraction

In the previous stage all the characters were reached and the beginning and end of each character and the space occupied by the character. At this stage the process of extracting attributes is carried out for the purpose of generating a series of observations, then

summoning a model. The hidden Markov is designed according to the character location, and the probability of the sequence of character views will then be calculated .And output the distinctive character. Repeat these steps on the rest of the letters sequentially.

6-1-4 Features vector

The vector of attributes consists of (8) eight variables, each variable representing an attribute that we previously found
vector elements

a) Number of transitions Horizontal (0-1)
b) Number of points
c) Curved direction
d) Character location
e) Number of optical dots Black over the line Basis
f) Number of transitions Vertical (0-1)
g) The existence of the point
h) The presence of the loop

Number of optical dots Black over the line Basis
Number of transitions Horizontal (0-1)
Number of transitions Vertical (0-1)
Number of points
The existence of the point
Curved direction
The presence of the loop
Character location

Figure (5): Vector features and elements

6-1-5 Represent elements of the hidden Markov model

Four models were designed to distinguish the printed Arabic text by the location of the letter in the word (primary, intermediate, final، or separate).
The Markov sample type used is a left-to-right parallel mode.

6-1-6 Elements of the hidden Markov model designed for the initial character

Elements of Markov's hidden model of the primary character are as follows:

First, the possibility of distribution of the primary case is the possibility of the occurrence of the situation

Mi when $i=1,2,\dots\ 9$ In time (t) it is placed in a vector (A) be dimensions 1 X N where (N) Represents the number of cases, N=9.

A= [1.0 0 0 0 0 0 0 0 0]

Second: Matrix probability of transition between previous cases in the vector (A) size = N X N

According to the proposed model was its size (9 X 9).

The table shows the values of the vector (A).

Table (1): the probability of transition between cases of the primary character

State	M1	M2	M3	M4	M5	M6	M7	M8	M9
M1	0	0.3636	0.6365	0	0	0	0	0	0
M2	0	0	0	0.5	0.5	0	0	0	0
M3	0	0	0	0.6430	0.3570	0	0	0	0
M4	0	0	0	0	0	1.0	0	0	0
M5	0	0	0	0	0	0	0.2222	0.2222	0.5557
M6	0	0	0	0	0	0	0	0.3077	0.6923
M7	0	0	0	0	0	0	0		1.0
M8	0	0	0	0	0	0	0		1.0
M9	0	0	0	0	0	0	0		1.0

Followed by a probability distribution account for the path of view codes for the same characters in their other locations where a matrix of observations is called B This matrix consists of N x M, N=9 and M=17 Where (N) represents the number of instances of Hidden Markov model (HMM) designer, and (M) represents the number of expected view codes in each case, As shown in the table (2) below.

Table (2): Probability distribution of views - primary character

State	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
M1	1.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M2	0	0.375	0.375	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0
M3	0		0.6430	0.143	0.0713	0.6430	0	0	0	0	0	0	0	0	0	0	0
M4	0	0	0	0	0	0	0.7692	0.2308	0	0	0	0	0	0	0	0	0
M5	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0
M6	0	0	0	0	0	0	0	0	0	0.6154	0.2308	0.1538	0	0	0	0	0
M7	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0
M8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0
M9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

In the same way we design the identification of the middle and final character, separated by Hidden Markov Model

And calculate the form elements

First: Calculate the probability of distribution of the primary case

Second: the matrix (A) probability of transition between cases
Third: The probability distribution matrix (B) for the observations symbols.

6-1-7 Probability calculation

After completing the matrices and testing all the letters are obtained Views and consist of a series of views per character by location in the word, (forward algorithm) was applied to calculate the probability of views, where the (forward algorithm). At the expense of the probability of observations by collecting the potential paths of all hidden situations that can be produced. For all characters at O 'sequential observations' and applied to training data to calculate the probability of sequential observations, the existence of models designed by the location of the letter with in the word. Two tables were created for all the characters for calculated probability the series of views and according to his position in the word. We include table (3) of the initial letter.

Character	Calculated probability of initial character
بـ	**0.007235288619995**
حـ	**0.0230809432983398**
سـ	**0.07792207792208**

Table (3): A sample of the results of the implementation of the front-end algorithm for three primary characters

Conclusions

Through this work we can conclude the following:

1. The hidden Markov model is designed to recognize fast performance and high accuracy as shown by Keep track of the system's performance in recognizing the word and the portal script line described in the chart.
2. The ability of the front-end algorithm to recognize the image of the entered character after converting it to a series of observations. Through calculations and output of the character whose results are compared to the input letter.

The Future Work

There are several suggestions for improving the performance of this system:
It is possible to develop a system that recognizes the written Arabic text of different types and sizes of lines

1. Develop the system to work on the separation of Arabic words in the pages containing images and geometric shapes
2. Develop the system in order to identify the Arabic text written by hand
3. Develop the system to identify the Arabic text written by hand.
4. The cutting stage causes some errors in the word We need to develop the system to overcome the cutting stage

5. 6. Development of the system to include identification of Hamza (ء) as well as identification of overlapping characters such as (لا ، أ ، آ ، ء ، لإ ، لأ ، إ)

References

١. بكر عبد الله خورشيد وآخرون: ٢٠١١حوسبة الحرف العربي حرف الميم المعزولة بخط النّسخ أنموذجا
٢. عمر ديدوح: ٢٠٠٧مقاربة توصيفية للتعرّف الآلي على الخط العربي اليدوي ، مجلة الآدا ب واللغات-جامعة قاصدي مرباح - الجزائر - العدد السادس -
٣. عاصم عبد الفتاح نبويّ، صبري عبد الله محمود٢٠٠١: تمييز حروف اللّغة العربية المكتوبة آليا باستخدام الشبكات العصبية ذات الانتشار الرجوعي
٤.عجرش، آمال سفيح ، "٢٠٠٥استخدام المنطق المضبب آلية لتمييز الحروف العربية"، رسالة ماجستير ، غير **منشورة**، قسم علوم الحاسبات، كلية العلوم، جامعة البصرة، العراق.
5. Sharma Amit Kumar and kishor Mr.R Rama, (2007) "pattern recognition: Different available approaches", proceeding of National conference on challenges & opportunities in information technology (COIT-2007) RIMT-IET, Mandi Gobindrh. www.rimtengg.com/coit 2007/.../coitindex.html
6. Jain Anil K., Duin Robert P.W. and Mao Jain chang,(2000) "statistical pattern recognition: A review", IEEE Transaction pattern analysis and Machine intelligence, ,.22.1.
7. Jannoud, Ismael Ahmed, (2007) "Automatic Arabic Handwritten Text Recognition System", American Journal of Applied sciences, 4(11): 857-864, 1546-9239.
8. Jurafsky Daniel and Martin James H, (2000) "speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition", 2nd Ed., prentice-Hall, ISBN: 0-13-095069-6, 2006.
9. Rabiner Lawrence R., (1989) "A Tutorial on Hidden Markov Models and selected Applications in speech recognition", proceedings of the IEEE, ,.77.
١٠. الكيم، سلوان تحسين فالح٢٠١١) ،" تصميم نظام لتمييز الحروف العربية باستخدام الخوارزميات.الجينية رسالة ماجستير غير منشورة، قسم علوم الحاسبات، كلية العلوم، جامعة البصرة.
١١.الكسو، ابتهاج عبد الحميد محمد ٢٠٠٥" استخدام الشبكات العصبية في تقدير رتب سلاسل ماركوف مع التطبيق على سلسلة جبل بطمة في محافظة **نينوى**"، أطروحة دكتوراه غير منشورة، قسم الإحصاء، كلية علوم الحاسبات والرياضيات، جامعة الموصل، العراق .
12. Sofia, Fatin Basher Abdul Ahead,(2003) "An Implementation of Arabic speech recognition", Unpublished Ph.D. Thesis, Department of mathematical science, college of computer and mathematical science, university of Mosul, IRAQ , .
13. Aazami, Farshideh Einsele,(2008) "Recognition of ultra-low resolution, Antialiased text with small font sizes", Unpublished Ph.D. thesis, Scientarium informaticarum, Faculty of science, University of Fribourg, Switzerland..
14. Dunham, Margaret H, (2002) "Data Mining introductory and advanced Topics", prentice Hall.
15.Li xiaolin, Parizeau Marc and plamondon Rejean,(2000) "Training Hidden Markov Models with multiple observations-A combinational Method",IEEE Transactions on PAMI,.PAMI-22,.4,.371-377,.
16. Attaluri, srilatha,(2007) "Detecting Meta Morphic Viruses using profile Hidden Markov Models", Unpublished M.Sc. thesis, computers science, thefaculty of the department of computer science, university of San Jose State,.
17. Jecheva Vaselina,(2006) "A bout some Application of Hidden markov Model in intrusion detection system", International conference and computer systems and Technologies-compsys tech'06,.
18. Khorsheed M.S,(2003) "Recognizing handwritten Arabic Manuscripts using a single Hidden markov Model", Pattern Recognition letters 24.,.