International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

The Application of Hidden Markov Models (HMMs) in Offline Arabic Handwritten Recognition

Rawia Ahmed[1], Mohammed Musa[2]

[1, 2]College of Computer Science and Information Technology, Sudan University of Science and Technology

Abstract: ***This paper aimed to introduce the fundamental issues of a Hidden Markov Model (HMMs). It attempts to discuss the: mathematical notation of HMMS, the three fundamental problems related to HMMs, and the topologies of HMMS. A further review of research done by HMMS has been discussed to show its application on offline Arabic handwritten.***

Keywords: HMM, OCR, Handwriting recognition

1. Introduction

Nowadays, many studies have been concerned with the Handwriting recognition. Researches on offline handwritten Arabic text (words, characters, sub-character) recognition have received more attention, because of the need to Arabic document digitalization.

Handwriting recognition is the task of transforming a language represented in its spatial form of graphical marks into its symbolic representation [1].

Over the time, many methods have been explored by researchers to recognize the Arabic handwriting [2]. HMM method became the popular one that used for recognition purpose. Beside text recognition, HMM has been used successfully to model many applications. These application may be speech, phoneme recognition, offline and online signature and protein domain identification, etc [3]. Due to successful implementation of HMM in speech recognition, it was applied in many studies for offline Arabic handwritten recognition, especially in limited dataset, because it gives robust results [4].

Hidden Markov Models (HMMs) is a statistical method that uses probability measures to model sequential data represented by sequence of observation vectors. The theory of Hidden Markov Models (HMMs) was first introduced by Baum et al[5]. It provides a powerful statistical framework for solving several applications. It offers several advantages for handwriting recognition, these advantages are: 1) Preserving time and correctness of text since segmentation is not required in HMM. 2) Availability of several free HMM tools. 3) Automated algorithms exist for training the HHM models. 4) The theory behind the Hidden Markov Model method is straightforward and easy to understand. 5) Features selection are language independent, in other words the same features can be used for different languages.

Rabiner, L.R defined HMM as a finite state machine having fixed number of states. The states of the model cannot be observed directly (hidden), only the output symbols of the state can be observed. It can be classified as discrete or continuous HMMs according to the output symbol [6]. In discrete HMMs every state have its own discrete probability distribution for each sample, the outputs may be characters from the dataset or vectors from a codebook. In continuous HMMs the emission probability distribution for symbols is continuous in each state and can be represented by a Gaussian mixture model [7].

Most information given in this paper is taken from Rabiner[6,8], for more details and tutorials readers can refer to these sources.

The next section gives a brief overview of HMMs elements. In Section (3), the three basic problems of HMMs are described. Section (4) covers the topologies of HMMs. The applications of HMMs in offline Arabic handwritten recognition are discussed in Section (5) and conclusion is presented in Section (6).

2. Elements of Hidden Marcov Model

According to Rabiner, L.R [6] notations, HMM can be defined by the following elements or characteristic:

- T = length of the observation sequence.
- N = number of states in the model
- M = number of distinct observation symbols per state
- $S = \{S_1, S_2, \dots, S_N\}$ distinct states of the Markov process
- Q_t = the state at time t
- $V = \{v_1, v_2, \dots, v_M\}$ set of possible observations symbols
- $A = \{ a_{ij} \}$ state transition probabilities where
  $a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, $1 \le i, j \le N$ ……{1}
- $B = \{b_j(k)\}$ observation symbols probability in state j where
  $b_j(k) = P(v_k \text{ at } \mid q_t = S_j)$, $1 \le i \le N, 1 \le k \le M$ ……{2}
- $\pi = \{\pi_i\}$ initial state distribution where
  $\pi_i = \{P(q_1 = S_i)\}$,, $1 \le i \le N$ ……{3}

From the above equations, the HHM can be used to give the following observation sequence
$O = (O_1, O_2, \dots, O_i)$, $O_i \in V, 1 \le i \le T$ ……{4}
As it can be noticed, HHM model depends on A, B,π matrices. Therefore, the HMM represents by λ parameter, where $\lambda = (A, B, \pi)$.

3. The Three Problems of HMMs

There are three basic problems needed to be solved in HMMs to be implemented in real world applications. These problems are: evaluation, optimal state sequence and training. The three problems and their solutions will be discussed in this section.

3.1 The evaluation problem

The evaluation is the first problem in the HMMs. It focuses on how to compute the probability that the observed sequence $P(O|\lambda)$ was produced by the given model $\lambda = (A, B, \pi)$, or in other words given the observation sequence $O = (O_1, O_2, \dots, O_T)$, and a model $\lambda = (A, B, \pi)$ how do we efficiently compute the probability of $P(O|\lambda)$, which represents the probability of the observation sequence, given the model.

The most suitable solution to this problem is enumerating every possible state sequence of length T to do this, considering the following equations:

$$Q = (q_1, q_2, \dots, q_T) \quad \dots\dots\dots\dots\dots\dots\dots\dots\{5\}$$

where q_1 is an initial state then

$$P(O|Q,\lambda) = b_{q_1}(O_1).b_{q_2}(O_2)\dots b_{q_T}(O_T) \quad \dots\dots\dots\{6\}$$

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O/Q,\lambda) \quad \dots\dots\dots\dots\dots\dots\dots\{7\}$$

The probability of a state sequence Q can be written as

$$P(Q, \lambda) = \pi_{q_1}.a_{q_1q_2}\dots a_{q_{T-1}q_T} \quad \dots\dots\dots\dots\dots\{8\}$$

We can obtain the probability of $P(O|Q,\lambda)$ (the probability that O an Q occur concurrently) by producing the above two equations as follows:

$$P(O,Q|\lambda) = P(O|Q,\lambda).P(O,\lambda) \quad \dots\dots\dots\dots\dots\{9\}$$

The probability of O is obtained by the join probability over all possible state sequences of q.

$$P(O,\lambda) = \sum_{all\ Q} P(O|Q,\lambda).P(O,\lambda) \quad \dots\dots\dots\dots\dots\{10\}$$

$$= \sum_{q_1,q_2,\dots,q_T} \pi_{q_1}.b_{q_1}(O_1).a_{q_1q_2}b_{q_2}(O_2)\dots a_{q_{T-1}q_T}b_{q_T}(O_T)$$

What worth noticing, is that the above-reviewed equations are complex. They requires a lot of calculations, it involved $2T.N^T$ calculations (there are N^T possible state sequences and 2T calculations required for each state sequence). So, to solve this problem, the forward procedure can be used to reduce the equation to N2.T calculations [3].

3.2 The optimal state sequence problem

The second problem is focusing on finding the optimal state sequence associated of a given observation sequence (how to discover the hidden state of the model), on the other word if the observation sequence $O = (O_1, O_2, \dots, O_T)$ and λ model are given, then how do we choose a corresponding state sequence $Q = (q_1, q_2, \dots, q_T)$ that is optimal in some sense. Evaluation problem have specific solution while optimum solution is chosen out of multi-solution. Thus, it is difficult due to the definition of the optimal state sequence. There are several possible optimum solutions to this problem, out of which, the commonest solution is Viterbi algorithm, which finds best sequence for the given observation sequence. The theoretical of Viterbi algorithm can be found in[5],[10].

3.3 The training problem

The third problem of HMMs, is finding the appropriate method for adjusting the parameters (A, B, π) to maximize the probability of the observation sequence given the model P(O/λ).The Baum-Welch algorithm is an attractive method that used the forward and backward procedure to adjust the model parameters (A, B, π) to maximize $P(O|\lambda)$ [5].

4. The Topologies Of HMMs

HMMs can be classified according to the structure of the transition matrix to three categories. These categories are: fully connected HMM (ergodic) model, left–to–right HMM model and hybrid HMM model.

4.1 A fully Connected HMMs

In this model, every state should be accessible from every other state of the model in a finite numbers of steps. It must satisfy the following two constrains[3]:

- all a_{ij} coefficient is positive, where a_{ij}'s represents the elements of transition matrix.
- There is no distinguishing between starting and terminating states; every state can be accessible from every other state in the model.

This type is useful in speak application, because they satisfy the conditions. Figure 1.a illustrates a fully connected HMM with three states.

4.2 Left to Right HMMs

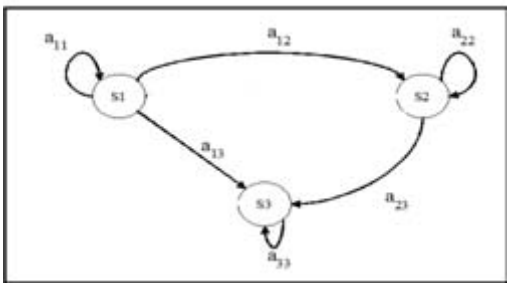
The fully connected model is not applicable for all applications. In some cases, the left to right model can be used. The key idea of this model is that the state grows from left to right with exception to the loops. Thus, the left to right model must satisfy the following constrains [6]:

- No transaction allowed to states with indexes lower than the current state index.
- The large jump from one state to other state is not allowed.
- The model begins from start state with lower index and ended with terminating state which has high index.

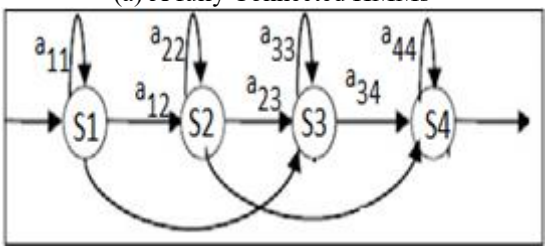
Figure1b: illustrates a left to right model with four states.

4.3 Hybrid HMMs

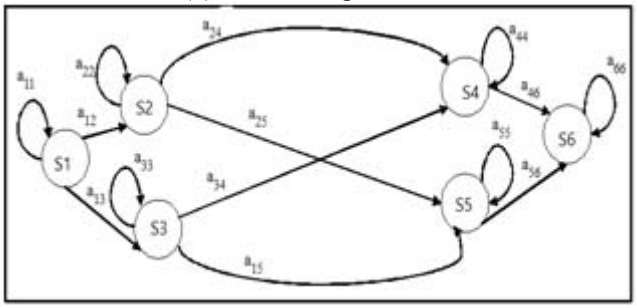
The combinations between the two topologies can be allowed in some practical cases for example Figure1c shows across joined of two parallel left to right model (four states in each model). They constructed new model which satisfied constrains of fully connected model and left to right model.



(a) A fully Connected HMMs



(b) A left to Right HMMs



(c) A hybrid HMMS

Figure 1: The three Topologies of HMMs

5. Applications of HMMs in Offline Arabic Handwritten Recognition

There are many studies reviewed the handwritten text recognition (offline, online). They were applied on different languages such as Arabic, Latin, Chinese, Hindi ..etc. HMM has been successfully applied in several large scale dataset for offline Arabic handwritten. It has advantages overall other recognition methods. The following are some of the works that were reported[11],[12],[13].

M. Dehghan et al.[14] used a discrete HMM with right to left topology to design A holistic system for the recognition of handwritten Farsi/Arabic words. The chain code directions histogram of the image has been extracted by a sliding window to represent the feature vectors. They constructed HMM for each word. Each model trained by Baum-Welch algorithm (60% of dataset). To improve the recognition rate the probability distributions of trained HMMs has been smoothing by SOFM codebook. They achieved better recognition rate in top10(69.47% before smoothing and 91.35% after smoothing when smoothing factor equal 0.001)

Mario Pechwitz and Volker Maergner[15] presented a semi continuous one-dimensional HMM system. The system designed to recognize Arabic handwritten words (26459 words from IFN/ENIT). The system consists of three steps. In the first step, skew, height, length, and baseline were normalized; baseline is normalized by using projection methods. Features are extracted by sliding window with three columns in the second step. On the final step a semi continuous HMM models were constructed per character. Each model has seven states and three transitions. Standard Viterbi Algorithm is implemented for training and recognition. The overall recognition rate reached to 89%.

M.S. Khorsheed [16] tried to overcome the overlapping problem in cursive Arabic script by designing a single HMM. This model consists of multiple models where each one built for representing one character. For example, we need four models to represent one model for the word"محمد" . The model depends on structural features extracted from the manuscript words. Features are extracted after implementing Zhang Suen thinning algorithm. The feature vectors are converted to discrete symbols by Vector Quantization algorithm. As mentioned previously, the system is single global model. It is created from ergodic character models. Each path through the global model represents a sequence of character which represents the desired word. The system tested on sample consists of 405 Arabic manuscript characters and achieved 97% (in top 5) accuracy rate.

ABDALLAH BENOUARETH et al.[17] proposed HMM system to recognize offline Arabic words. The system implemented on IFN/ENIT benchmark database. It based on discrete HMM with explicit state duration. After doing some preprocessing operations (baseline detection and thinning), they extracted the features by performing special segmentation methods to segment the word into frames. Then, feature vectors are constructed, and they were combination between statistical and structural features(41 feature). Vector Quantization algorithm is used to map the continuous features vector to discrete features. Right to left HMM topology with three transitions is used to recognize the character (one HMM for each word character). Because they used explicit state duration, the states of HMM are varied according to character length. This technique is useful and increases the recognition rate. The models were trained and tested by standard Viterbi algorithm. Results showed HMMs with explicit state duration (with Gamma distribution) improved the accuracy rate.

Ramy El-Hajj et al[18] attempted to solve the overlapping problem in Arabic character. They assume that some characters have ascending and descending strokes. These strokes may be overlapped with two neighboring characters. They improved their previous system[19] by adding extra HMM models to represent the contextual character models. Therefore, the HMMs are increased. Thus, the system has one dimension HMM for each Arabic character and contextual character models. They depends on lower and upper baselines features. The system tested on IFN/ENIT dataset and found that the contextual character models improved the recognition rate about 0.6 %.

6. Conclusion

We reviewed the important issues of HMMs. HMMs may be continuous or discrete or hybrid. Discrete HMMs are more applicable. Three problems need to be solved in HMMs, these problems are: how to compute the probability of the observation sequence, how to find the optimal state sequence associated with a given observation sequence and how to find the appropriate method for adjusting the HMM parameters to maximize the probability of the observation

sequence given the model. Several algorithms and methods are found to solve these problems.

References

[1] Plamondon, R. and Srihari, S.N. "On-line and Offline Handwriting Recognition: A Comprehensive Survey" . *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2000. 22(1): 63 – 84.

[2] Märgner, V. and H.E. Abed. *"Arabic handwriting recognition competition. in Document Analysis and Recognition",*. Ninth International Conference . 2007. IEEE

[3] Amin, A.," *Off-line Arabic character recognition: the state of the art".* Pattern recognition, 1998. **31**(5): 517-530

[4] Bunke, H., Wang, P. S. P. and Baird, H. S. (eds.), "*Hand Book of Character Recognition and Document Image Analysis*", World Scientific, 1994.

[5] Baum, L.E., et al., "*A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains".* The annals of mathematical statistics, 1970: 164-171.

[6] Rabiner, L.R., "*A tutorial on hidden Markov models and selected applications in speech recognition".* Proceedings of the IEEE, 1989. **77**(2): 257-286.

[7] Petrushin, V.A. "*Hidden markov models: Fundamentals and applications".* in *Online Symposium for Electronics Engineer*. 2000.

[8] Rabiner, L. and B.-H. Juang, "*An introduction to hidden Markov models".* ASSP Magazine, IEEE, 1986. **3**(1): p. 4-16.

[9] Forney Jr, G.D., "*The viterbi algorithm"* Proceedings of the IEEE, 1973. **61**(3): p. 268-278.

[10] Alma'adeed, S., C. Higgins, and D. Elliman, "*Off-line recognition of handwritten Arabic words using multiple hidden Markov models".* Knowledge-Based Systems, 2004. **17**(2): p. 75-79.

[11] Kessentini, Y., T. Paquet, and A.B. Hamadou, "*Off-line handwritten word recognition using multi-stream hidden Markov models".* Pattern Recognition Letters, 2010. **31**(1): p. 60-70.

[12] Marti, U.-V. and H. Bunke, "*Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system".* International journal of Pattern Recognition and Artificial intelligence, 2001. **15**(01): p. 65-90.

[13] Dehghan, M., et al., *Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM".* Pattern Recognition, 2001. **34**(5): p. 1057-1065.

[14] Märgner, V., H. El Abed, and M. Pechwitz. "*Offline handwritten arabic word recognition using hmm-a character based approach without explicit segmentation".* in *Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document*. 2006. SDN06.

[15] Khorsheed, M.S., "*Recognising handwritten Arabic manuscripts using a single hidden Markov model".* Pattern Recognition Letters, 2003. **24**(14): p. 2235-2242.

[16] Benouareth, A., A. Ennaji, and M. Sellami. "*HMMs with explicit state duration applied to handwritten Arabic word recognition*. in *Pattern Recognition", 2006. ICPR 2006. 18th International Conference on*. 2006. IEEE.

[17] El-Hajj, R., C. Mokbel, and L. Likforman-Sulem. "*Recognition of Arabic handwritten words using contextual character models"*. in *Electronic Imaging 2008*. 2008. International Society for Optics and Photonics.

[18] El-Hajj, R., L. Likforman-Sulem, and C. Mokbel. "*Arabic handwriting recognition using baseline dependant features and hidden Markov modeling"*. in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. 2005. IEEE.

Author Profile

Rawia Ibrahim Omer Ahmed: Received the B.Sc from University of Khartyoum, received the M.E degree in Computer Science from Faculty of Mathematical Science, Khartoum University – Khartoum, Sudan in 2003..Presently pursuing P.D degree in Computer Science from College of Computer Science and Information Technology, Sudan University of Science and Technology – Khartoum, Sudan.

Prof.Mohamed Elhafiz Mustafa Musa : Received the P.D degree in Computer Science from College of Computer Engineering Department, Faculty of Engineering, Middle East Technical University – Ankara, Turkey in 2003. He is associated currently with the Computer Science and Information Technology, Sudan University of Science and Technology – Khartoum, Sudan.