International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184, Volume 8, Number 1 (February 2019), pp.1-8

http://www.meacse.org/ijcar

Towards Offline Arabic Handwritten Character Recognition Based on Unsupervised Machine Learning Methods: A Perspective Study

Ahmad Hasasneh[1],Nael Salman[2] and Derar Eleyan[2,3],
[1]College of IT, Palestine Ahliya University, Palestine
[2]College of Engineering,Palestine Technical University-Kadoorie, Palestine
[3]College of IT, Birzeit University, Palestine

Abstract

This paper proposes an alternative approach for the problem of Arabic handwritten character recognition. The proposed model is based on Deep Belief Networks (DBNs) which are unsupervised machine learning methods. A greedy layer-wise fashion based on Restricted Boltzmann Machines and contrastive divergence learning algorithm will be used to train such model. Previous studies have shown that DBNs are capable to extract a set of sparse features, which can be used to code the initial data in an efficient way. The assumption is that such representation must improve the linear separation among the different classes and thus a simple classification algorithm, like softmax regression, should be sufficient to achieve accurate recognition rates. The literature reviewed showed that this alternative approach has not been considered yet in the context of Arabic character recognition, which deserves to be investigated and evaluate its performance for such problem.

Keywords: Arabic Character Recognition, Restricted Boltzmann Machines, Contrastive Divergence, Deep Belief Networks, Sparse Features, SoftmaxRegression

Introduction

Character recognition (CR) has recently become one of the most important tasks in pattern recognition and artificial intelligence. The process of CR usually depends on several factors like various shapes, various font sizes, noise, and broken lines or characters etc. and thus it influences the recognition results. Despite recognizing letters which are inconsistent in shape can be a difficult task; it however depends on machine learning algorithms and methods used to represent the actual characters in the feature space.

Usually, the process of character recognition can be achieved through two main phases: detection and classification. In the first phase, each given image is pre-processed, improved, and then the regions of interest (ROI) is segmented based on the characters' attributes such as shape and colors. The output of the segmented regions should represent the possible characters in the given image. Indeed, accuracy and speed of detection play an important role in obtaining accurate and fast recognition process. In the recognition or classification phase, a set of features (patterns) for each segmented character is first extracted and then used to classify and recognize the character. These features can be used as a reference to understand the differences among the classes.

The remainder of the paper is organized as follows. Section 2 describes the characteristics of Arabic characters. Section 3 presents the current approaches that have been proposed to investigate the problem of

Arabic character recognition (ArCR). The proposed model for recognition of Arabic handwritten characters is presented in section 4. Finally, conclusions and future works are presented in section 5.

Characteristics of Arabic Characters

Some languages like Persian, Ordo and other share common features with Arabic characters. So, researchers working on ArCR can lend their propagate results to these languages. However, it has been shown that the developed Latin CR techniques are unfortunately inappropriate for the problem of ArCR [1]. Arabic machine-printed and handwritten character recognition is a complex task and presents several challenges to obtain accurate results [2]. Some of these challenges are due to the facts that some of the Arabic characters have strong similarities and correlations. Also, the number of characters in Arabic language is 28, written from right to left not like the Latin language. The Arabic language has special characteristics and rules different from other languages, for instance, the letter may have four different shapes, depending on the letter position in each word including: (initial, medial, final and isolated). In other words, depending on the character's connection to the preceding and subsequent characters as shown in Table I. Also, an Arabic letter might use a combination of dots (1-3) or short hand symbols above and below these shapes (e.g. HAMZA) as illustrated in Table I. These rules and special characters complicate the process of recognition and thus a sophisticated recognizer is needed. So, instead of dealing with 60 separate classes, in this work, we are initially interested in investigating a new machine learning approach that could improve the feature space representation of the primary Arabic Handwritten Alphanumeric letters and then simplify the later classification tasks. So that, the dots or secondary symbols are not considered yet in this research proposal, leading to a total of 24 separate characters and digits need to be classified and recognized.

Table 1.Shapes of Arabic characters depending on their position in the word

Final	Mid	Initial	Isolated
ـا			أ
ـب	ـبـ	بـ	ب
ـت	ـتـ	تـ	ت
ـث	ـثـ	ثـ	ث
ـج	ـجـ	جـ	ج
ـح	ـحـ	حـ	ح
ـخ	ـخـ	خـ	خ
ـد	ـد	د	د
ـذ	ـذ	ذ	ذ
ـر	ـر	ر	ر
ـز	ـز	ز	ز
ـس	ـسـ	سـ	س
ـش	ـشـ	شـ	ش
ـص	ـصـ	صـ	ص
ـض	ـضـ	ضـ	ض
ـط	ـطـ	طـ	ط
ـظ	ـظـ	ظـ	ظ
ـع	ـعـ	عـ	ع
ـغ	ـغـ	غـ	غ
ـف	ـفـ	فـ	ف
ـق	ـقـ	قـ	ق
ـك	ـكـ	كـ	ك

ل	ـل	لـ	ل
م	ـم	مـ	م
ن	ـن	نـ	ن
ه	ـه	هـ	ه
و	ـو	و	و
ي	ـي	يـ	ي

While most of the previous ArCR works focused on machine-printed characters, some recent attempts brought attention towards handwritten Arabic Characters [3]–[8]. Despite the existence of these recent attempts that have been proposed to address the problem of ArCR [3]–[7], [9]–[11], the advanced studies have been conducted on other languages, like Latin language. However, the current works can generally be categorized into offline and online CR systems. Concerning offline ArCR proposed models, an offline handwritten Arabic text recognition is proposed in [6] based on Hidden Markov Models and re-ranking, where the authors have used the IFN/ENIT database which contains 32,492 words, to conduct extensive experiments and obtain accurate results. Gheith et al. [4] proposed another offline handwritten Arabic character recognition based on complex pre-processing, multiple features extraction algorithms and selection tools, and sophisticated classification techniques. In particular, after pre-segmenting the letters into main body and secondary components, they extracted the moment features from the whole letter, its main body, and its secondary components. They further selected the efficient features based on multi-objective genetic algorithm [4]. Finally, they used a Support Vector Machine (SVM) to evaluate the classification error for the extracted efficient features [4]. Abandah et al [5] used a more complicated algorithm to achieve offline handwritten Arabic letter recognition. The proposed approach starts by extracting 96 features from the letter's secondary components, main body, skeleton, and boundary. Varying sizes of the extracted features are then selected using five feature selection techniques. Although this approach obtained the highest recognition rates, its time complexity however is very high. The authors in [9] developed another model based on the properties of Bezier´ curves to achieve multi-font Arabic characters' recognition. Mirza [12] developed a database of Arabic characters of different font sizes and introduced a recognition algorithm based on Minimum Distance Classifier. Testing the algorithm on a set of samples obtained recognition rate of more than 97% for two different font sizes, it however excluded the dotted characters [12]. Moreover, recent approaches based on very deep neural networks (DNNs) [8], [13] have addressed the problems of image recognition and ArCR respectively. In particular, the proposed model in [8] enhanced the classification accuracy rate and resulted in reduction of the overall classification complexity compared to the previous presented works. In a recent survey of Optical Arabic Character Recognition (OACR) Systems, Alhomed and Jambi [14] provided a review of several attempts towards OACR and presented the similarities and differences of these attempts. They also proposed suggestions to encounter challenges related to Font-size, Font-type, Video caption, and Low resolution of captured Arabic characters' images [14]. Saudagar and Mohammed [11] developed OACR algorithm based on Zhang-Suen thinning method and then proposed traversing extraction algorithm. After implementing and testing the proposed algorithm, the obtained results, a recognition rate of 98.1%, outperformed the existing attempts of OACR approached [11].

On the other hand, several recent works have been developed to achieve online Arabic characters' recognition, a review of the recent attempts for online Arabic characters recognition can be found in [6]. Some of them proposed to use Bayes classification method [10]. This method computes the tangent differences using histograms and then finds Gibbs modeling of the class-conditional probability density functions. Furthermore, the authors in [7] proposed another online Arabic characters' recognition system based on neural networks. Although the proposed approach in [7] works well with some letters, the efficiency of the algorithms degrades for other characters.

After presenting some of the current proposed methods of ArCR, we have seen that most of them are based on supervised machine learning methods, i.e. they require labeling the data to perform the training task.

Achieving a reasonable performance of the recognition rate therefore requires increasing the amount of labeled data for the training phase. In addition to that, these empirical hand-crafted feature detectors required to use sophisticated recognizers, like SVM, to reach reasonable accuracy of recognition. On the other hand, we have also noticed that most of the existing approaches used multiple pre-processing techniques to achieve promising results. Increasing the amount of training data and using complex and sophisticated filters will result in complicating the overall recognition algorithm. Moreover, most of the presented methods are based on two main phases of image coding followed by the classification process. Therefore, the problem of ArCR is still an open question and needs to be tested using alternative approaches. In other words, ArCR requires a new machine learning approach that leads to an appropriate code in the feature space and thus simplifies the later processes with accurate results. To achieve that, Deep Belief Networks (DBNs) have recently emerged and used in different applications [15]. In contrast to the current empirical methods, DBNs introduce nonlinearities in the coding scheme and exhibit multiple layers of Restricted Boltzmann Machines (RBMs) as shown in Figure 2 (right). So that the lower RBM layer in a DBN becomes a visible layer to the higher one and so on. Each RBM layer can be trained using a Contrastive Divergence (CD), which is a fast learning technique first proposed by Hinton [16]. It has also been shown that DBNs are capable to learn localized, sparse, and efficient features from a database of tiny images [17]. A sparse code means that a small number of neurons can be sufficient for coding an image. Besides, it makes the images linearly separable after being coded using the extracted features. Consequently, a complex classification task is transformed into an easier one and thus a softmax regression, which is a simple linear classifier, can be used to perform the classification process in a fast and accurate manner [18]. As mentioned before, these recent methods coupled with tiny images have been successfully investigated and used to achieve various recognition tasks, for instance semantic place recognition and robotics [18], [19], gestures recognition [20], phone and speech recognitions [21], brain artifacts signals classification [22], and traffic sign recognition [23]. However, to our knowledge these recent approaches have not been considered yet in the context of ArCR.

Proposed Model

The fundamental purpose of this work is to propose an alternative, simple, and robust approach for Arabic handwritten characters recognition as stated before. As shown in Figure 1, this approach involves four main phases including: data collection, data pre-processing and normalization, features extraction and image coding, and finally character recognition.

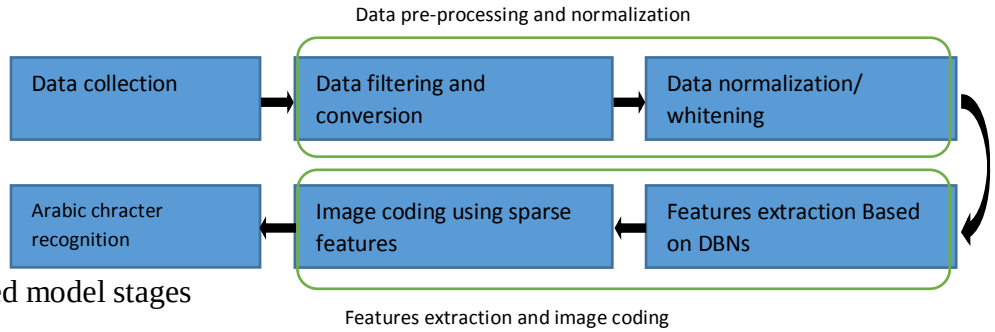

Fig. 1. Proposed model stages

A. Database Description

In previous works, researchers used various datasets to test their ArCR models. For convenient comparison with the existing approaches, we plan to use two databases which are: ADBase [24] (a dataset of 10 Arabic digits from zero to ten) and HACDB [25] (a dataset of the handwritten Arabic letters). These databases have been successfully used in the context of ArCR based on DNNs [8], or based on Fuzzy Turning Function [26], or based on SVM [27]. A clear description of the two databases can be found in [8].

B. Data Pre-processing and Normalization

To achieve accurate results to the problem of Arabic characters' recognition, we have seen that most of the current approaches used various image pre-processing and filtering techniques before performing the learning and recognition tasks. Data pre-processing aims to improve the quality of letter's image by highlighting the useful information and hidden details within an image, like letter's edges and corners, and eliminating the noises and unnecessary information. To achieve that, smoothing and sharpening filters will be applied on gray images.

On the other hand, it has been shown that natural images, for instance images of handwritten Arabic characters, have strong correlations, similarities, and statistical redundancies [2]. Removing these similarities and correlations require data normalization or whitening technique, which is a major requirement for extracting localized sparse features [28]. Data whitening removes the first and second order statistics presented in the initial data. The theoretical background of data whitening is based on the fact that the data is first centered, projected onto their principal axes through computing the eigenvectors of the variance-covariance matrix, and then divided by the variance of each image. On the other hand, data normalization simply corresponds to a local brightness and contrast normalization. It locally normalizes the data by simply subtracting the mean of a given image and then dividing the result by the standard deviation of the image elements.

C. Features Extraction and Characters Recognition

After appropriate pre-processing, a feature learning algorithm is needed to extract a set of sparse features which can be used later to efficiently code the images in the feature space. To achieve that, this paper proposes to use new unsupervised machine learning models; DBNs. These generative methods constructed of multiple RBMs layers of binary and Gaussian units. These units correspond to the hidden layer or features detectors. Since data is already normalized, the input layer to the first RBM corresponds to zero-mean Gaussian activation values and is often used to compute the first hidden layer as shown in Figure 2 (left). Then, the model is reconstructed by using the hidden units to reconstruct the visible units and finally recompute the hidden layer using the reconstructed visible layer. As shown in Figure 2 (left), there are symmetric and undirected connections between the visible and hidden layers. These connections represent the weight matrix or features need to be learned after the RBM network is converged to the right solution. The learning process is based on minimizing the energy function according to the quality of the reconstructed image and by using the contrastive divergence algorithm. Of course there are several parameters, like the learning rate, momentum, and weight decay, play important roles in extracting interesting features, so fine tuning these parameters is important to make them appropriate to the given ArCR datasets. More theoretical information of the proposed model, its training methodology, its learning parameters, and some of its applications can be found in [18]–[20], [23], [29].

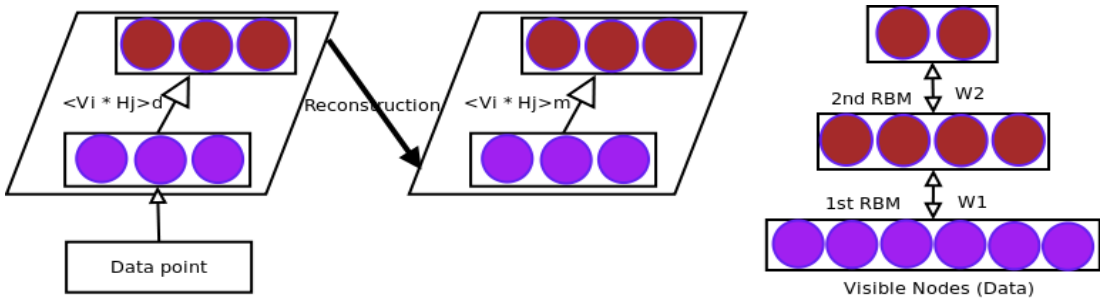

Fig. 2.Left: the RBM training methodology based on a contrastive divergence learning algorithm, where Vi and Hj represent the visible and hidden units respectively and Wij represents the connection weight matrix or actual features need to be extracted. Right: a DBN with two RBM layers.

It has been shown that using binary units for the reconstructed visible layer is not appropriate for multivalued inputs like pixel levels of handwritten Arabic letters. A Gaussian-Bernoulli RBM will thus be used to train the network as suggested by Hinton [29]. Since DBNs are composed of multiple RBMs layers as shown in Figure 2 (right); one RBM layer can therefore be trained at a time (layer-wise training methodology) using Contrastive Divergence learning algorithm as shown in Figure 2 (left). The number of RBM layers and the size of each RBM layer (number of units) depend on the final recognition results and on the overall classification complexity. Finally, assuming DBN computes a linear separable signature of the initial data, a softmax regression in the feature space must be sufficient for the Arabic character recognition purpose. One can see that DBNs also contribute in reducing the dimensionality of the network and thus speeding-up the classification process. In other words, using such code in the feature space simplifies the classification task instead of using sophisticated classifiers, like SVM, in the input space. Meanwhile, the classification task will also be investigated using SVM. In case of similar results; this will underline that the images become linearly separable after coding them using the sparse features learned by DBN.

Conclusions and Future Works

OACR for typed and handwritten characters is an emerging field that still needs more academic research and industry support to reach the level attained in other languages like English, French…etc. Having effective OACR systems will definitely help in preserving and archiving scripts typed or written Arabic on digital platforms and make them available for faster and simpler access. Several attempts introducing different approaches towards tackling OACR are available in the literature, which are based on complex techniques. In this paper, we illustrated how DBNs coupled with small images, 32x32 pixels for instance, can be effectively used as alternative and simpler approach in the context of ArCR. The hope is that such an approach can achieve comparable results to those recent works proposed [8], [11], [12], [14], [26], [27].

Our future works are to empirically implement and investigate this approach for the printed and handwritten Arabic characters and digits. In particular, after features extraction, the images will be sparsely coded and used to perform the classification process. Assuming such code in the feature space becomes linearly separable, which should be gained by DBN, a simple classifier, like softmax regression, will thus be used to classify and recognize the Arabic characters. As mentioned earlier, a sophisticated classification technique, like SVM, will also be used to investigate the linear separation capability which it should be acquired by the DBN. On the other hand, we will investigate the effect of normalization and whitening techniques on feature extraction and evaluate their influences on the final classification results. Further testing and validation is still needed to consider cases of scripts containing dotted characters, secondary symbols and characters of different font-types and font-size.

Acknowledgment

The authors would like to acknowledge Palestine Technical University - Kadooriefor funding the conference participation and publication of this article.

References

[1]H. Aljuaid, D. Mohamad, M. S.-I. J. of Computer, and undefined 2011, “Evaluation approach of Arabic character recognition,” igi-global.com.

[2] H. Barlow, “Redundancy reduction revisited,” Netw. Comput. Neural Syst., vol. 12, no. 3, pp. 241–253, Jan. 2001.

[3]G. Abandah, T. M.-D. E. S. Journal, and undefined 2010, “Feature selection for recognizing handwritten Arabic letters,” researchgate.net.

[4]G. Abandah, N. A.-J. of C. Science, and undefined 2009, "Novel moment features extraction for recognizing handwritten Arabic letters," researchgate.net.
[5]M. Kherallah, F. Bouri, A. A.-E. A. of Artificial, and undefined 2009, "On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm," Elsevier.
[6]J. AlKhateeb, J. Ren, J. Jiang, H. A.-M.-P. R. Letters, and undefined 2011, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking," Elsevier.
[7]K. Addakiri, M. B.-I. J. of Computer, and undefined 2012, "On-line handwritten arabic character recognition using artificial neural network," researchgate.net.
[8]M. Mudhsh, R. A.- Information, and undefined 2017, "Arabic Handwritten Alphanumeric Character Recognition Using Very Deep Neural Network," mdpi.com.
[9]AzzedineMazroui and AissaKerkourElmiad, "BÉZIER CURVES TO RECOGNIZE MULTI-FONT ARABIC ISOLATED CHARACTERS."
[10]N. Mezghani, A. Mitiche, M. C.-I. T. on, and undefined 2008, "Bayes classification of online arabic characters by gibbsmodeling of class conditional densities," ieeexplore.ieee.org.
[11]A. Saudagar, H. M.-I. A. J. I. Technol., and undefined 2018, "Arabic character extraction and recognition using traversing approach.," iajit.org.
[12]N. M.-I. J. of Science and undefined 2018, "Printed Arabic Characters Recognition Based on Minimum Distance Classifier Technique," scbaghdad.edu.iq.
[13]K. Simonyan, A. Z. preprint arXiv:1409.1556, and undefined 2014, "Very deep convolutional networks for large-scale image recognition," arxiv.org.
[14]L. Alhomed and K. Jambi, "A Survey on the Existing Arabic Optical Character Recognition and Future Trends," researchgate.net.
[15]G. E. Hinton, S. Osindero, and Y.-W.Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[16]G. E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," Neural Comput., vol. 14, no. 8, pp. 1771–1800, Aug. 2002.
[17]B. Olshausen, D. F.-C.opinion in neurobiology, and undefined 2004, "Sparse coding of sensory inputs," Elsevier.
[18]A. Hasasneh, E. Frenoux, P. T.-I. (2), and undefined 2012, "Semantic Place Recognition based on Deep Belief Networks and Tiny Images.," researchgate.net.
[19]A. Hasasneh, "Robot semantic place recognition based on deep belief networks and a direct use of tiny images2012 ",.
[20]A. Hasasneh, S. T.-G. J. of Computer, and undefined 2017, "Towards Arabic Alphabet and Numbers Sign Language Recognition," computerresearch.org.
[21]G. Dahl, A. Mohamed, G. H.-A.in neural information, and undefined 2010, "Phone recognition with the mean-covariance restricted Boltzmann machine," papers.nips.cc.
[22]A. Hasasneh, N. Kampel, P. Sripad, … N. S.-J. of, and undefined 2018, "Deep Learning Approach for Automatic Classification of Ocular and Cardiac Artifacts in MEG Data," hindawi.com.
[23]A. Hasasneh, Y. Daraghmi, and N. Hasasneh, "Towards Accurate Real-Time Traffic Sign Recognition," ijcit.com.
[24]E. E.-S. SherifAbdelazeem, "AHDBase." [Online]. Available: http://datacenter.aucegypt.edu/shazeem/. [Accessed: 17-Dec-2018].
[25]A. Lawgali, M. A.-… P. (EUVIP), undefined 2013, and undefined 2013, "HACDB: Handwritten Arabic characters database for automatic character recognition," ieeexplore.ieee.org.
[26]F. Zaiz, M. Babahenini, A. D.-E. A. of Artificial, and undefined 2016, "Puzzle based system for improving Arabic handwriting recognition," Elsevier.
[27]M. Elleuch, N. Tagougui, and M. Kherallah, "Towards Unsupervised Learning for Arabic Handwritten Recognition Using Deep Architectures," 2015, pp. 363–372.

[28]A. Coates, A. Ng, … H. L. the fourteenth international conference on, and undefined 2011, "An analysis of single-layer networks in unsupervised feature learning," jmlr.org.
[29]G. E. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," 2012, pp. 599–619.