Engineering Science and Technology, an International Journal xxx (2016) xxx–xxx

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Full Length Article

Recognition-based online Kurdish character recognition using hidden Markov model and harmony search

Rina D. Zarro *, Mardin A. Anwer

Software Engineering Department, Salahaddin University-Erbil, Erbil, Kurdistan, Iraq

ARTICLE INFO

Article history:
Received 1 August 2016
Revised 4 November 2016
Accepted 18 November 2016
Available online xxxx

Keywords:
Character recognition
Evolutionary computation
Kurdish character recognition
Hidden markov model
Harmony search

ABSTRACT

In this paper a hidden Markov model and harmony search algorithms are combined for writer independent online Kurdish character recognition. The Markov model is integrated as an intermediate group classifier instead of a main character classifier/recognizer as in most of previous works. Markov model is used to classify each group of characters, according to their forms, into smaller sub groups based on common directional feature vector. This process reduced the processing time taken by the later recognition stage. The small number of candidate characters are then processed by harmony search recognizer. The harmony search recognizer uses a dominant and common movement pattern as a fitness function. The objective function is used to minimize the matching score according to the fitness function criteria and according to the least score for each segmented group of characters. Then, the system displays the generated word which has the lowest score from the generated character combinations. The system was tested on a dataset of 4500 words structured with 21,234 characters in different positions or forms (isolated, start, middle and end). The system scored 93.52% successful recognition rate with an average of 500 ms. The system showed a high improvement in recognition rate when compared to similar systems that use HMM as its main recognizer.

1. Introduction

Nowadays, the growth of portable and mobile technologies, lead to a necessity to develop a character or text recognition application, as most of the current devices does not integrate a keyboard. However, character recognition systems represent a complex and wide research area. Developing a successful system requires a tremendous dataset to be analyzed which takes a long time. The main procedures that effects the recognition success rate may resides in the methods used in character preprocessing, extracted feature and recognition. However, the challenges of developing an efficient system are not limited to its recognition rate and accuracy but also its recognition time especially in online recognition [23]. Therefore, it is very common that a recognition systems may compromise the success rate in order to achieve an acceptable recognition time.

Kurdish, as many languages which use Urdu, Arabic and Persian based characters or alphabets, has gained a wide research interest in character recognition for the past few years. The research motivation in Kurdish character recognition originates from its special and complicated styles of writing related to the same languages. Complications in Kurdish recognition come from the possibility of writing a single word with one stroke or with many strokes, depending on the user writing style. Moreover, it is possible to write a single character in various styles. Besides, the word may change when diacritics (dots, hamza, . . .etc) are written.

In this paper, a combination of Hidden Markov Model (HMM) and harmony search (HS) is used to recognize Kurdish characters extracted from word segmentation process. The character is preprocessed to create a more consistent movement pattern, and also removing effects such as incomplete chain, hocks. HMM is used as first stage recognition step for extracting the possible group of characters the character may belong. Then, HS is applied to identify the closest fit character from the group when compared to a set of dataset. A special matching criteria is used in HS fitness function.

The rest of this paper is organized as follows. In Section 2 a review of similar works in the field of online Kurdish characters recognition is introduced. Section 3 introduces the Kurdish character systems with explanation to the character's structure and the problems and challenges exists in the system. Section 4 explains the structure of the online Kurdish character recognition system.

* Corresponding author.
E-mail addresses: rina.zarro@yahoo.com (R.D. Zarro), mardinsherwany77@gmail.com (M.A. Anwer).

Peer review under responsibility of Karabuk University.

http://dx.doi.org/10.1016/j.jestch.2016.11.016
2215-0986/

Please cite this article in press as: R.D. Zarro, M.A. Anwer, Recognition-based online Kurdish character recognition using hidden Markov model and harmony search, Eng. Sci. Tech., Int. J. (2016), http://dx.doi.org/10.1016/j.jestch.2016.11.016

It explains the preprocessing phase where characters are enhanced before features are extracted for the recognition phase. The recognition phase explains the HMM classifier and HS recognizer structures. The result is introduced in Section 5. Section 6 presents the last drawn conclusion.

2. Related works

Kurdish character recognition systems can be categorized into two groups: offline and online [23]. Offline systems deal with machine printed or handwritten text or characters. These data are processed more as images and pixels to extract as much features and information from these images as possible. The time may not be an important or relevant factor compared to output text recognition accuracy. On the other hand, online systems deal with direct coordinate data obtained from writing strokes made on the tablet device. Since preprocessing and information extraction in online systems must occurs in real time, a fast algorithm should be integrated in the system. Therefore, it may produce lower recognition accuracy compared to offline as the processing time is reduced.

Although, there are not many researches that deals with Kurdish handwritings [33], there are various studies which handles this field but with similar set of characters such as Arabic, Urdu and Persian. These languages have many characters in common with Kurdish alphabets. Therefore, the paper will focus on these languages for problem presentation and literature review, as well as, comparison. Research in Kurdish like characters was initiated in the end of 80s. The earliest systems were only handling optical recognition for scanned documents in the form of printed or handwritten. The online systems were basically dealing with isolated Arabic characters, without taking character forms in cursive writings under consideration for keeping the system as simple as possible [15]. Later, advance methods in online Kurdish like character systems were introduced with most of them handling recognition of isolated characters [1,16]. These systems were modified and enhanced by the end of the 90s to deal with segmented or segmentation free text based recognition system [5].

Hidden Markov model or briefly HMM is a statistical tool used to model a sequence of events that can be characterized by a Markov process [34] it was widely used effectively in speech or character recognition systems for different languages such as Latin, Korean, Kurdish and many others [14,4,10,25]. In languages having similar alphabet set as Kurdish, HMM was integrated as a standalone recognizer or a stage in the recognition process. The significance between these works is either the extracted character features, or character encoding type used in training of HMM. Daifallah et al. [12] used a HMM for Arabic script recognition, in which, the segmented characters are treated as images. These images are preprocessed and scaled to extract characters' image seven momentums. These momentums are used to train and classify the result for the HMM recognizer. Rashwan et al. [35] used a feature vector in HMM training and recognition. The system uses a sliding window through the word to generate a set of feature per word. The used vector consists of lossless differential luminosity coding based features. Additionally, a dynamic range normalization parameters estimator was used to detect the effective dynamic range of the features to calculate the normalization parameters from the population of the feature vectors in the training data. Biadsy et al. [8] used a feature vector and HMM to recognize Arabic words. The feature vector consists of three main features which are local angle, super segment and loop feature. These feature was quantized to create a dataset for the purpose of HMM training. Naz et al. [30] presented a full review on using HMM algorithm for handwriting recognition for various languages which share the same alphabets. The review included HMM algorithm use for recognizing Arabic, Persian, Urdu abd Pishtu handwritings. The review explore various feature extraction techniques and classification.

HMM was combined with other algorithms in order to enhance the feature extraction results. Amor et al. [6] used a Hough transformation and HMM for multi-font Kurdish character recognition. The Hough transform was used to extract unique and meaningful features within the character in which a set of lines were extracted. These lines were used to train the HMM. The HMM for each character had 4–7 left to right states in which each state return to itself or go to the next state. The system was intended to recognize Kurdish printed characters with different fonts. This approach was later used by the same authors to develop a more robust features based on Wavelet transformation [7]. The System uses both Hough features and Wavelet features to train HMM for a better results. The basic results obtained from these HMM recognition based system was acceptable. However, the result obtained was always limited and there was only a slight opportunity to modify for obtaining better results. Razzak et al. [36] used a fuzzy set theory with HMM for script based Urdu characters. The HMM input consists of fuzzy rules created to identify characters within the script. This approach modified the result from 81% to 87% successful recognition rate.

Evolutionary or swarmed based algorithms represented a generic population based metaheuristic optimization algorithms. It depends on generating a new population from the existence one following some fitness function. Initially, these algorithms were successfully used in offline character recognition rather than online recognition because it takes a longer time to find a feasible solution. This short time recognition characteristic is crucial in online systems. Genetic algorithm (GA) was initially used to optimize feature selection problem. The GA generated population was used to find the smallest feature subset from a wider feature range which optimize the separation between different classes. The algorithm was tested successfully on digit dataset [11] and Persian [40]. In addition to feature selection, GA was used in offline character recognition system as well. It was successfully applied to Latin [27] and Arabic [2,28]. In addition to GA, other heuristic methods were employed in character recognition based on swarm optimization. Nebti et al. [31] applied particle swarm optimization (PSO) and a combination of back propagation neural network and bee colony for digit recognition based on image momentum calculations. Particle swarm was used as a statistical classifier for comparing the generated feature with the digit data set feature to obtain the optimal class while the second method was determining the assigned class using back propagation to classify the digits and in case of no classification is obtained the bee colony was used to assign the class. Sarfraz et al. [37] used PSO with moment invariants for Arabic character recognition. PSO applied to optimize the weight given to each feature in the feature vector so that it can maximize the possibility to find the right class. Singh and Shrivastava [39] conducted a performance evaluation of feed forward Neural Network with three different soft computing techniques on handwritten English alphabets. The study concluded that using Neural Network with evolution algorithm gave a better recognition accuracy than using stand-alone Neural Network. The proposed study observed that, there are more than one converge weight matrix in character recognition for every training set. Finally, Singh et al. [38] integrated GA and feed forward NN for evaluating the recognition of Hindi curved scripts. GA is used to make the search process more efficient to determine the optimal weight vectors from the generated population. The study analyzed that the proposed method of a descent gradient of distributed error with the GA known as hybrid distributed evolutionary technique for the multilayer feed forward neural performs better in terms of accu-

Please cite this article in press as: R.D. Zarro, M.A. Anwer, Recognition-based online Kurdish character recognition using hidden Markov model and harmony search, Eng. Sci. Tech., Int. J. (2016), http://dx.doi.org/10.1016/j.jestch.2016.11.016