Research Journal of Applied Sciences, Engineering and Technology 8(19): 2016-2025, 2014
DOI:10.19026/rjaset.8.1193
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: December 09, 2013 Accepted: June 08, 2014 Published: November 20, 2014

Research Article

The Optical Character Recognition for Cursive Script Using HMM: A Review

[1, 3]Saeeda Naz, [1]Arif I. Umar, [1]Syed H. Shirazi, [2]Muhammad M. Ajmal and [2]Salahuddin
[1]Department of Information Technology, Hazara University, Mansehra, Pakistan
[2]COMSATS Institute of Information Technology,
[3]GGPGC No.1, Department of Higher Education, KPK, Abbottabad, Pakistan

Abstract: Automatic Character Recognition has wide variety of applications such as automatic postal mail sorting, number plate recognition and automatic form of reader and entering text from PDA's etc. Cursive script's Automatic Character Recognition is a complex process facing unique issues unlike other scripts. Many solutions have been proposed in the literature to solve complexities of cursive scripts character recognition. This paper present a comprehensive literature review of the Optical Character Recognition (OCR) for off-line and on-line character recognition for Urdu, Arabic and Persian languages, based on Hidden Markov Model (HMM). We surveyed all most all significant approaches proposed and concluded future directions of OCR for cursive languages.

Keywords: Character, hidden Markov model, ligature, optical character recognition

INTRODUCTION

Optical Character Recognition (OCR) converts text images into text file. The main objective of OCR is to mimic the reading ability of human being with accuracy and high speed. Its applications include the following among many others; machine reading for handicapped persons, preservation and online access of historical documents and advanced scanning.

The printed OCR systems for Latin, Japanese and Chinese Languages have attained maturity. Several commercial products of OCR for these languages are available. The handwritten OCRs are far behind from the maturity level. There is a huge demand for improved products of handwritten OCRs. Printed noisy text images and text images captured by camera also require attention of the researchers in the field of English OCR. The cursive languages are spoken, written and understood across the globe. These languages among others includes; Arabic, Persian (Farsi), Urdu, Uyghur, Jawi, Pashtu and Sindhi. The historical literature of these language contain great academic treasure in several domains like poetry, history, spiritualism, astrology, medicines and mystics. Each of the cursive languages has its own set of alphabets, vocabulary, grammar rules, fonts and writing style. The popular fonts styles are; Naskh, Riq'a, Nasta'liq, Thuluth, Kofi and Diwani. The most popular and common writing styles are Naskh and Nasta'liq. Urdu is generally written in Nastaliq. Arabic and Persian use Naskh fonts for writing. Few commercial products of Printed Arabic OCR in rudimentary forms are available. The cursive nature and forms of letter depending on its position to create words are creating challenges for researcher in the segmentation stage of the character recognition. Urdu language unlike Arabic language has some peculiarities due to more alphabet/letters and some unique properties. These peculiarities make OCR in Urdu language more complex and challenging. But the tradeoff between the efforts of producing printed OCR for Urdu and its potential befits urge us to devote our energies to this significant domain of the research. It will empower automatic conversion of ancient Urdu, Persian and Arabic documents into printed books. This will result in the availability of huge ancient oriental academic treasurer on the web, which is now available in printed form in libraries. Despite of its significance, currently no commercial OCR product for Urdu language is available. Many researchers are investigating the complexity and challenges offered by the Urdu language to produce an optimal OCR product for Urdu language.

Cursive languages: Cursive script languages are written cursively from right to left. The letters are joined with neighbour letters usually within word or sub-word. The shape of letters and whether joined or standalone depends on their property (Naz *et al*., 2013a). Logically, a letter could have one shape or two to four different shapes depending on its position in the sub-word or word. A letter is in connected sequence or in isolated form could appear at the start of the word, at middle of word or at end of the word. This is called joiners. However, some letters have two forms namely final and isolated. It may join the letter which it follows

Corresponding Author: Saeeda Naz, Department of Information Technology, Hazara University, 21300, Mansehra, Pakistan

She is a Dr. Noor
وہ ایک ڈاکٹر نور ہے
وهي الدكتور نور
او نور یک دکتر است

Fig. 1: The sentence “She is a Dr. Noor” in Urdu, Arabic, and Farsi, respectively

but do not join the letter that follows it. These are called non-joiners.

For example first letter in Noor is the joiner and therefore it has four shapes. The second and the last letters in Noor are non-joiners and have two forms as shown in Fig. 1. The shape of these letters depends on the context. A sentence in Urdu, Persian and Arabic is shown in Fig. 1 to understand the writing styles of the languages.

Basic terminologies: The commonly used terminologies in OCR are following:

Word definition: It is composition of one or more than one ligature e.g., the word Ahmad "احمد" have 2 ligatures and Hamad "حمد" has one ligature.

Primary or main ligature definition: The primary ligature is the longest continuous portion of the character that is written before lifting the pen. It is a character which is unique combination of more than one letters. It is also called PWs (Pieces of Words) and main strokes in some papers (Husain *et al*., 2007) e.g., there are 3 ligatures in ”پشاور ” i.e., ” و”,” پشا”, and ” ر ” and one ligature in " جشن "

Secondary ligature definition: The secondary ligature is a set of diacritics, dots that are written up or bellow after the main ligature.

Primitives definition: It is the basic geometrical shape.

Strokes definition: These are the basic sub-pattern that cannot segment more.

Optical character recognition (OCR): On the basis of input mode, OCR is classified as offline and online. The block diagram of a character recognition system is shown in Figure 2. The offline OCR deals with the image of the already written text --- handwritten and machine printed --- and its input is acquired by means of an optical scanner or digital camera. In contrast, in the online OCR, the input text is obtained directly by means of tablet, a PDA, or a stylus. The online character recognition is perhaps easier than its offline counterpart because extra information, such as stroke coordinates time information and handwriting style of the user, is available. A typical OCR system may include some or all of the five mechanisms, namely the:

- Image acquisition
- Pre-processing

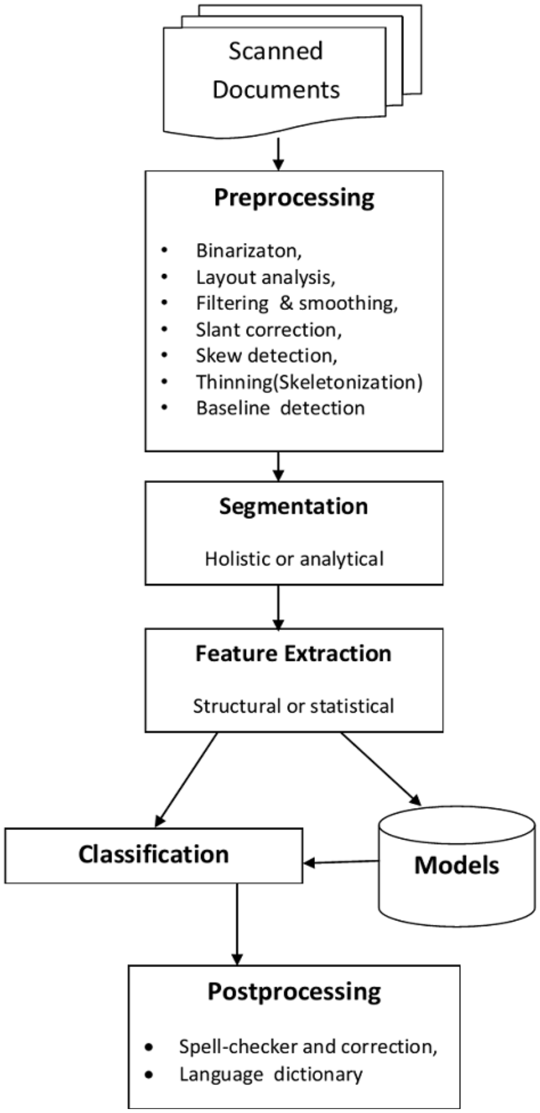

Fig. 2: Steps of OCR

- Segmentation
- Feature extraction
- Recognition/classification (Al-Badr and Mahmoud, 1995) followed by some post-processing.

The rest of the paper will present a comprehensive literature review of the OCR methods proposed based on HMM.

ISSUES IN CURSIVE SCRIPT LANGUAGES

The literature reflects that Arabic language out of the prevailed cursive script languages has been favorite choice of the researchers for construction of OCR. Persian language is the next. Urdu being a popular language of the glob have not yet get the require attention of the researcher consequently not a single OCR is materialized for it. Some serious effort have been made in this direction for Urdu and Jawi (Spoken in Malaysia) unlike other languages like Pashtu and Sindhi. Urdu scripts are complex and more challenging because of; large set of characters, similar multi-shaped characters, context sensitivity and position of character (Razzak *et al*., 2009; Satti and Saleem, 2012). The Nasta'liq font style adds further to challenges because language is written diagonally with no fixed baseline, no standards for slopes, context sensitivity caused by filled or false loops and character/ligature overlaps (Slimane *et al*., 2012).

The Nastaliq font style poses some extra challenges as compared to the Naskh style, these are:

- Diagonality is introduced by Nasta'liq writing style in Urdu that makes this language more complex for researcher in the field of OCR.
- Text written in Nasta'liq takes less space horizontally due to diagonality introduced by the Nasta'liq, which adds complexity in segmentation and recognition of the text.
- The filled loops in Wao, Fay, Meem and Khaaf create confusion in recognition of these letters after thining process.
- The number and position of dot with respect to primary ligature also add to challenges.
- The intra-ligature and inter ligature overlapping in Urdu text being Arabic based text add to challenges in the segmentation and recognition (Naz 2013).
- A baseline is a virtual line on which the text is joined (Husain *et al*., 2007; Naz *et al*., 2013b). Nastaliq style unlike the Naskh style has multiple baselines due to digonality nature of Nasta'liq that makes baseline a difficult task.

Generally, the projection profile technique is utilized for Arabic script, which calculates the horizontal profile (called histogram in some texts) of text lines (due to the latter's bulky interline spacing) and segments where the profile has zero values. Yet, this technique is not applicable to Nasta'liq, where the ligatures overlap in horizontal/vertical projections and display minor spacing among the lines.

HMM BASED CURSIVE SCRIPT CHARACTER RECOGNITION

In the literature, Hidden Markov Models (HMMs) have been used to recognize various types of observation sequences including speech, word spotting, on-line handwriting and off-line handwriting in document images. We describe the predominant application of HMMs given by segmentation-free and segmentation based recognition of cursive script for off-line and on-line handwritten ligatures, words or text lines. For HMM, the scanned images of paragraphs will be segmented into text lines or words and the ground values/labels are written as a segmentation cue point for letters in segmentation based recognition system. In segmentation free approach, text line will separated into ligatures or sub-words and code book will be generated for each ligature groups and then subjected to number of steps of OCR. The former approach uses the HMM encoder and the latter approach employs the HMM evaluator, (Fig. 3).

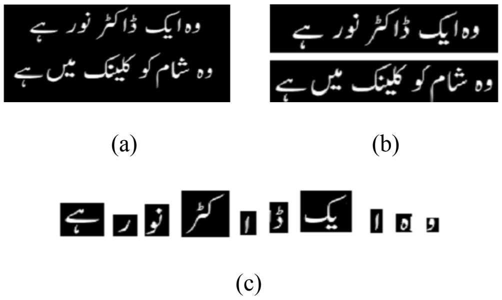

Fig. 3: Examples from Urdu, (a) A simple paragraph, (b) segmentation of paragraph into textlines, (c) Text line segmentation into words

A Hidden Markov Model (HMM) is used by (Javed 2007) for classification of segmented stokes of the ligature by calculating the DCTs as features for improvement of performance of recognition. In another effort, HMM toll kit (HTK) is trained as a classifier for the recognition of Nastaliq font style (Javed *et al*., 2010). A post processing step is employed which consequently significantly enhanced the accuracy of classifier results. For testing, 3,655 ligatures are used from 5,000 common words and 3,375 ligatures are accurately identified. The authors claimed 92% recognition accuracy.

The method specified in Akram *et al*. (2010) is an extension of Javed, 2007 their scheme extracts Nasta'liq ligatures independent of the font size. The Splines method is implemented on the input image of the ligature, resulting in the outlines. To control the points in the splines and the input ligatures the outlines are then scaled and it is resized to train the OCR. The scaled outline is then converted to the image form so that the system could accomplish recognition. The proposed system was assessed on the Urdu single character ligatures and attained 98% accuracy rate for the manually generated data and a 96% accuracy rate for the data scanned from several books and magazines.

The BBN Byblos Pashtu OCR System (Decerboet *et al*., 2004) applied a script independent OCR by implementing fourteen-state HMMs for Pashtu language. This system is also tested successfully for English, Arabic and Chinese documents.

An extension of their previous work is suggested in (Razzak *et al*., 2010), which applies the fuzzy logic and HMM for on-line recognition of 1,800 ligatures for the Nasta'liq and Naskh fonts. For each stroke by the HHM the dataset is trained, while further classification is performed through the application of the fuzzy logic rules to the starting and ending shape of the characters. The proposed hybrid approach provides reasonable result for large deviation in handwritten strokes and declines the comparison and computation. Finally, the mapping of secondary strokes and analyzed primary strokes are done by using the fuzzy logic for the recognition of valid ligature. The authors reported 87.6% accuracy for the Nasta'liq font and 74.1% accuracy rate for the Naskh font.

A related on-line multi-font numeral recognition system is proposed by (Razzak *et al*., 2009). It depends on HMM, fuzzy rules and a combination of the two in the case of Arabic and Urdu. They explored the similarities and dissimilarities of the two languages to design both the on-line and off-line OCRs. All the three methods reported significantly accurate results. The proposed system is analyzed on 900 samples that were taken from 30 trained users. It demonstrated an accuracy of 97.4, 96.2 and 97.8% by each method respectively the authors reported that the proposed system has some unsolved issues caused by the complexities of the languages Ghods and Kabir (2010) used extracted features from online isolated Persian letters and classified these features using ID3. The features include horizontal and vertical directions, angles, number of strokes, x-y coordinates. Ghods *et al.* (2013a) in another work combined two separate horizontal and vertical trajectories for features extraction for on- line handwritten recognition. Two classifiers were trained on these trajectories. Finally, these results fused and used as an input to the HMM classifier. The consequence of delayed strokes such as the accents for building of Persian on-line words recognition with full ligatures for classification through HMM model was studies by the same authors. The delayed strokes concepts reduced the size of the lexicon. The proposed system was evaluated on "TMU-OFS" dataset consisting of 1000 Persian ligatures (2013b).

Discrete HMM and Kohonen self-organizing vector quantization methods have proposed for recognition of 17820 handwritten Farsi/Arabic words (Dehghan *et al*., 2001). The features selected by histogram of chain-code and sliding window and information of neighbourhood stored in the "Self-Organizing Feature Map" (SOFM). The recognition rate was obtained upto 65%.

In Sajedi *et al.* (2007), the Persian characters were divided in 18 groups using HMM and got 90.8% accuracy rate.

Aulama *et al.* (2011) proposed HMMs combined with the Viterbi algorithm for characters recognition at the sub-word level using extracted statistical information and estimated probabilistic parameters for HMM.

A system is implemented in two stages for recognition of Arabic words by Al-Hajj *et al.* (2007). First, three classifiers (HMM) using likelihood were employed with features based on pixels to find out the best top ten candidates. Second the sum rules, the vote rules and then neural network implemented on the classifiers' results for getting combined decision. The recognition rate was 90.96% using the handwritten IFN/ENIT database. These arrangements were used in (Mohamad *et al*., 2009) for performance improvement. Three homogenous classifiers using HMM was employed.

The recognition rate of proposed system on IFN/ENIT (v1.0p2) with lexicon (1,000 entries) was 90.26, 94.71 and 95.68% for top 1, top 2 and top 3.

Semi Continuous Hidden Markov Model (SCHMM) was employed in (Benouareth *et al*., 2006; Benouareth *et al*., 2008; Benouareth *et al*., 2006a) for recognition of Arabic words in unconstrained environment on handwritten IFN/ENIT data base using holistic approach for segmentation. The combination of structural and statistical features were mined using sliding window approach, which were operated according to uniform and non-uniform schemes (constant and variable width) of segmentation of images into vertical frames. Further, they considered morphological complexities of hand-written letters and for this they analysed minima and maxima of the vertical projection histogram. For uniform segmentation, 81.02 and 91.74% recognition rates were achieved for top 1 and top 10. For non-uniform segmentation, 83.79 and 92.12% recognition rates were achieved for top 1 and top 10. They also used and compared distributions of Gauss, Gamma and Poisson for the explicit state duration modelling and concluded that continuous distribution gave better results for Arabic character recognition than discrete distribution with non-uniform segmentation.

Another system based on HMM (semi-continuous 1-dimensional) in (El-Abed and Margner, 2007) used the similar strategy for Arabic handwritten words. In four directions of the word, the skeleton's length computed from five equal horizontal zones were used for statistical features extraction. The recognition rate was upto 89.1 and 96.4% for top 1 and top 10 candidates and evaluated on IFN/ENIT (v1.0p2) database.

A hybrid method of a HMM classifier followed by re-ranking trained on V1.0p2 (26,459 words written by 411 writers) and V2.0p1e (32,492 words by more than 1000 writers) of IFN/ENIT database in (AlKhateeb *et al*., 2011) for recognition Arabic handwritten word. Sliding window used for extraction of intensity features for HMM and structural/topological features were extracted for re-ranking. The proposed scheme achieved recognition rate 83.55% with fusion of re-ranking and 82.32% for top 1 on v2.0 p1e and 89.24% with re-ranking and 86.73% without for v1.0 p 2.

AlKhateeb *et al*. (2011a) presented Hidden Markov Models (HMMs) and Dynamic Bayesian Networks (DBNs) using DCT coefficients/mean values of the overlying blocks of the complete Arabic word for HMM and pixels of the entire Arabic word for handwritten Arabic words recognition on IFN/ENIT database consisting 32,492 Arabic words. They compared experimental results of HMM (nearly 83%) and DBN (66.56%) which showed that HMM outperformed than BDN.

Kundu *et al*. (2007) employed Variable Duration HMM (VDHMM) with two lexicons using 45 structural and statistical features. The authors evaluated the proposed system on IFN/ENIT database and testified 60% recognition rate. Hidden Markov Models (HMM) classifier proposed for off-line handwritten Arabic word Recognition without explicit segmentation on IFN/ENIT database (Margner *et al*., 2006). Sliding window and statistical Karhunen-Loeve Transform (KLT) were employed for feature selection and provide the recognition rate of 74.69%.

Multiple HMM scheme had presented by Hamdani *et al*. (2009) using both on-line and off-line features for off-line Arabic handwritten. In this scheme, Pixel values, Densities and Moment invariants, Pixel distribution and Concavities for off-line features and statistical method in (Benouareth *et al*., 2006) for On-line features and got 49.48-63.90% and 81.72% recognition rate by single classifier and multiple classifier, respectively.

Dreuw *et al*. (2009) used constrained maximum likelihood linear regression (CMLLR) transformed features for training the HMM and tested on IFN/ENIT database and displayed recognition accuracies of 94.18 and 88.78% for set-d (6735) and set-e (6033), respectively.

The multistream hidden Markov models with lexicon (2100 words) employed for recognition of off-line Arabic handwritten word through directional and colour density features without explicit segmentation and achieved an accuracy of 79.6% in (Kessentini *et al*., 2010). In Elbaati *et al*. (2009) narrated Arabic handwritten OCR using HMM. Beta-elliptical extractor and dimension of Arabic word/letters used for feature extraction and came out with recognition rate of 54.13%.

Ahmed and Azeem (2011) used ADAB Database for online Arabic text recognition sing HMM htk toolkit. They removed the delayed stokes and used directional, acceleration and delta features and claimed accuracy rate between 89.72% and 95.27% on different sets. In another effort Azeem and Ahmed (2012) combined the offline and online features for online Arabic text recognition using HMM htk toolkit.

In Azeem and Ahmed (2013) fixed the width of the input word to three pixels and also fixed the space among the various parts of the word to an experimental threshold in the pre-processing step, which improved 3.6% performance. The non-uniform segmentation adopted which also improved the performance up to 3.97%. The images divide into multi concavity layers with adding new concavity spaces and then concavity space and sliding window used to extract concavity and densities of the foreground pixel, respectively. Three HMM models used for slant correction and recognition off-line Arabic handwriting words. The proposed system reported 93.44% recognition rate on IFN/ENIT database with lexicon.

In 2008, Natarajan proposed a script-independent approach for multi-lingual BBN Byblos off-line handwriting recognition (OHR) based on 14-state left to right Hidden Markov Models (HMM) using percentiles, vertical and horizontal derivatives, angle and correlation features. For more detail of BBN Byblos OCR, refer to (Natarajan *et al*., 2001) for feature extraction and recognition. It was named as OHR. The OHR system experimented on 26,459 images of IFN/ENIT database consisting of 4 sets a, b, c and d. and obtained 89.4% recognition rate. The result of BBN Byblos OCR system, Natarajan *et al*. (2008) improved by incorporating the calculated baseline into percentile feature calculation and showed the improvement of 1% absolute gain and 3.1% relative gain in recognition accuracy on 1404 handwritten Arabic words (1389K training words and 15K testing words) by Natarajan *et al*. (2011). In this study, features like Percentiles of intensity values (Menasri *et al*., 2007) Angle, Correlation, Energy (PACE) and Gradient, Structure and Concavity (GSC), were extracted in (Xiang *et al*., 2012).

The connected component was employed using sliding window for training of HMM for recognition of different font/size/font and size of Arabic printed text having ultra-low resolution (Slimane *et al*., 2012). The authors reported recognition rate of 69.9% by global multi-font system for word and using cascading multi-font system, the average recognition rates was 93.7% for word on the Arabic Printed Text Image (APTI) database, (Slimane *et al*., 2009) consisting of 113284 distinct words and 45, 313, 600 total words.

Cao *et al*. (2012) demonstrated 14-state HMM based on different features which are image intensity percentile, local angle and correlation signifying the orientation of the stroke, frame energy, concavity, gradient and Gabor filter for Arabic handwritten, printed and mixed pages in The Word Error Rates of 26.4% for handwritten lines, 9.4% for typed written lines and 20.2% for mixed lines depending on the 25736 transcript pages with dictionary (600K words).

Awaida and Khorsheed (2012) demonstrated HMM classifier based on Run-Length Encoding (RLE)

Table 1: Summarize the cursive script' OCR system using HMM / HMM htk toolkit

Authors	Language	Features	Classification	Dataset	Accuracies
Javed *et al*. (2010)	Urdu	DCT	HMM htk toolkit	1259 Unique ligatures from 5000	92%
Akram *et al*. (2010)	Urdu	DCT	HMM htk toolkit	569 Manually generated samples and 1260 scanned samples	98% and 96\% (1 character ligatures), 89% (2 characters ligatures)
Decerboet *et al*. (2004)	Pushtu	Statistical	HMM htk toolkit	27000 characters	1.6% CER and 5.1% WER
Razzak *et al*. (2010)	Urdu	Structural and statistical	HMM and Fuzzy	1800 Nasta'liq ligatures	87.6% (Nasta'liq) and 74.1% (Naskh)
Razzak *et al*. (2012)	Urdu	Fuzzy rules	HMM and Fuzzy	1800 Nasta'liq ligatures	89.4\%
			HMM and Fuzzy	1800 Nasta'liq ligatures	
Razzak *et al*. (2009a)	Urdu	Fuzzy rules	Fuzzy, HMM and Hybrid	900 Ligatures	97.4%, 97.8% and 96.2%
Dehghan *et al*. (2001)	Farsi	Histogram of chain-code and sliding window	Discrete HMM and Kohonen self-organizing vector quantization	17820	65%
Ghods and Kabir (2010)	Online isolated Farsi letters	Structural (direction, angles, number of strokes etc.)	ID3	4000	94-99%
Ghods *et al*. (2013b)	Online Farsi ligatures	Fusion of statistical and structural	HMM	1000 Online ligature samples	Testing: 87.5% Training: 92.9%
Ghods *et al*. (2013a)	Farsi	Structural	HMM with lexicon reduction,	1000 (1200 groups)	Top1: 85.2% Top10:96.7%
Margner *et al*. (2006)	Arabic	Sliding window and statistical (KLT)	HMM with lexicon	32492	74.69%
Kundu *et al*. (2007)	Arabic	Structural and statistical	Variable length HMM with lexicon	32492	60%
Dreuw *et al*. (2009)	Arabic	CMLLR transform	HMM	6735 and 6033	94.18% & 88.78%
Elbaati *et al*. (2009)	Arabic	Directional and colour density by Beta-elliptical extractor	HMM	32492	54.13%
Kessentini *et al*. (2010)	Arabic	Directional and colour density	Multistream HMM with lexicon	32492	63.5%-70.5%
AlKhateeb *et al*. (2011)	Arabic	Sliding window and structural	HMM and re-ranking	26459 and 32492	83.55%
Hamdani *et al*. (2009)	Arabic	Densities and moment invariants, pixel values, pixel distribution and concavities, on-line features	Multiple HMM Single HMM	32492	81.93% 49.48%--63.90%
Benouareth *et al*. (2006)	Arabic	Statistical and structural	HMM	26459	88.12%
Benouareth *et al*. (2008)	Arabic	Pixels densities, concavities based on baselines and 9-structural based on skeleton	HMM	26459	83.79%
Al-Hajj *et al*. (2007)	Arabic	Pixels densities	Fusion of multiple HMM and neural network	26, 459	90.96%
Mohamad *et al*. (2009)	Arabic	16 Distribution features (Pixel densities) and 12 features representing concavity configuration with respect to baselines	Multiple homogeneous HMM	26, 459	Less than 91%

Table 1: Continue

Khorsheed (2007)	Arabic	Statistical features	HMM htk toolkit (mono-model) HMM htk toolkit (Tri-model)	16, 743	70.2% 85.9%
Natarajan *et al*. (2008)	Arabic	Percentiles, angle and correlation	14-states HMM	26, 459	89.4%
Natarajan *et al*. (2011)	Arabic	Baseline-dependent Percentiles, Angle, Correlation , Energy (PACE) and Gradient, Structure, Concavity (GSC)	14-states HMM	26, 459	90.4%
AlKhateeb *et al*. (2011a)	Arabic	DCT coefficients/mean values of the overlapping blocks of the whole Arabic word	HMM	32, 492	nearly 83%
Cao *et al*. (2012)	Arabic	Image intensity percentile, local angle and correlation representing the orientation of the stroke, frame energy, gradient, concavity, and Gabor filter	HMM	25736 transcriped pages	Word error rates of 26.4% for handwritten lines, 9.4% for typed written lines and 20.2% for mixed lines
Awaida and Khorsheed (2012)	Arabic	Run-length encoding (RLE)	HMM	94, 418	96.65%
Azeem and Ahmed (2012)	Arabic	Directional, aspect, number of foreground pixels and gradient	HMM	ADAB database	97.78%
Menasri *et al*. (2007)	Arabic	Baseline-dependan	HMM and NN	IFN/ENIT and	87%
Slimane *et al*. (2012)	Arabic	Connected component	HMM	45, 313, 600	69.9%-93.7%
Xiang *et al*. (2012)	Arabic	Densities of foreground, concavity and arc length	HMM	26459	97.99%

features, (Khorsheed and Al-Omari, 2011) using 94,418 words images from APTI (Arabic Printed Text Image) database. The achieved average recognition rate on the letter level was 96.65%. The same authors put forward HMM and nearest neighbour classifier along with generation of angle, distance, horizontal and vertical span features for recognition of independent writer off-line handwritten Arabic numeral in (Mahmoud, 2008). Using a database of 21,120 digits (2171 numbers), that has written by 44 writers. The average recognition rates were 97.99% and 94.35% using the HMM and the nearest neighbour classifiers, respectively. The misclassification was nearly 1% in the case of HMM.

Awaidah and Mahmoud (2009), presented scale invariant and translation invariant technique HMM classifiers for off-line handwritten Arabic numerals recognition with multi-resolution feature extraction methodology using GSC algorithm. The features were edge curvature in a neighbourhood of a pixel, short strokes types which span number of pixels and certain concavities that can span across the image. Different grid sizes have used for segmentation of an image. The technique uses a database of 21120 digits was used. This technique gave average of 1.01% higher recognition rate than the previous one with overall 99.0% recognition rate of accuracy. Azeem and Ahmed (2012) used off-line HMM to improve the results of HMM-based on-line Arabic handwriting recognition system presented by combining the efficient on-line features and off-line features (dividing the image into non uniform heights). The recognition rate was improved by 2.38%.

The HMM and NN classifiers were combined in (Menasri *et al*., 2007) using IFN/ENIT benchmark database based on explicit grapheme segmentation and Seventy-four baseline dependent features vectors. Hidden Morcov Model Toolkit (HTK) was used for recognition of Arabic handwritten text based on sliding window features like densities of foreground, concavity and arc length on databas of IFN/ENIT consisting of 26459 total words and achieved 85.43\% average recognition rate by the proposed system in (Xiang *et al*., 2012).

We summarized the different OCR systems for different languages using HMM with accuracies and datasets in Table 1.

CONCLUSION

The literature review concludes that for the Naskh writing styles the HMM model and HMM Took Kit (HTK) have performed very well in word/character recognition. But for the text line or word recognition in Nasta'liq writing style only few instances have been reported where in HMM and HMM Tool Kit (Htk) or variation of the tool kit have been explored. In future we need to investigate an optimal and sophisticated algorithm using HHM htk tool kit or it variations for Nastal'iq text line or word recognition.

REFERENCES

Akram, Q.U., S. Hussain and Z. Habib, 2010. Font size independent OCR for Noori Nastaleeq. Proceedings of Graduate Colloquium on Computer Sciences (GCCS), Vol. 1, NUCES Lahore.

Aulama, M.M., A.M. Natsheh, G.A. Abandah and M.M. Olama, 2011. Optical character recognition of handwritten Arabic using hidden Markov models. Proceedings of SPIE.

Al-Hajj, R., C. Mokbel and L. Likforman-Sulem, 2007. Combination of HMM-based classifiers for the recognition of Arabic handwritten words. Proceeding of IEEE 3th International Conference on Document Analysis and Recognition (ICDAR'07), 2: 959-963.

Azeem, S.A. and H. Ahmed, 2013. Effective technique for the recognition of off-line Arabic handwritten words using hidden Markov models. IJDAR'13, 16: 399-412.

AlKhateeb, J.H., J. Ren, J. Jiang and H. Al-Muhtaseb, 2011. Offline hand-written Arabic cursive text recognition using Hidden Markov Models and re-ranking. Pattern Recogn. Lett., 32(8):1081-1088.

AlKhateeb, J.H., O. Pauplin, J. Ren and J. Jiang, 2011a. Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition. Knowl. Based Syst., 24: 680-688.

Awaida, S.M. and M.S. Khorsheed, 2012. Developing discrete density hidden Markov models for Arabic printed text recognition. Proceeding of IEEE International Conference on Computational Intelligence and Cybernetics (CyberneticsCom'12), pp: 35-39.

Awaidah, S.M. and S.A. Mahmoud, 2009. A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. Signal Process., 89(6): 1176-1184.

Azeem, S.A. and H. Ahmed, 2012. Combining on-line and off-line systems for Arabic handwriting recognition. Proceeding of 21st IEEE International Conference on Pattern Recognition (ICPR'12), pp: 3725-3728.

Ahmed, H. and S.A. Azeem, 2011. On-line Arabic handwriting recognition system based on HMM. Proceeding of International Conference on Document Analysis and Recognition (ICDAR'11), pp: 1324-1328.

Al-Badr, B. and S.A. Mahmoud, 1995. Survey and bibliography of Arabic optical text recognition. Signal Process., 41(1): 49-77.

Benouareth, A., A. Ennaji and M. Sellami, 2006. HMMs with explicit state duration applied to handwritten Arabic word recognition. Proceeding of IEEE 18th International Conference on Pattern Recognition (ICPR'06), 2: 897-900.

Benouareth, A., A. Ennaji and M. Sellami, 2008. Semi-continuous HMMs with explicit state duration for uncon-strained Arabic word modeling and recognition. Pattern Recogn. Lett., 29(12): 1742-1752.

Benouareth, A., A. Ennaji and M. Sellami, 2006a. Semi-continuous HMMs with explicit state duration applied to Arabic handwritten word recognition. Proceeding of 10th International Workshop on Frontiers in Handwriting Recognition.

Cao, H., J. Chen, J. Devlin, R. Prasad and P. Natarajan, 2012. Docu-ment recognition and translation system for unconstrained Arabic documents. Proceeding of 21st International Conference on Pattern Recognition (ICPR'12).

Decerboet, M., E. MacRostie and P. Natarajan, 2004. The BBN Byblos Pashto OCR system. Proceedings of the 1st ACM Workshop on Hardcopy Document Processing, pp: 29-32.

Dehghan, M., K. Faez, M. Ahmadi and M. Shridhar, 2001. Handwritten Farsi (Arabic) word recognition: A holistic approach using discrete HMM pattern recognition. Elsevier, 34: 1057-1065

Dreuw, P., D. Rybach, C. Gollan and H. Ney, 2009. Writer adaptive training and writing variant model refinement for off-line Arabic handwriting recognition. Proceeding of IEEE 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp: 21-25.

Elbaati, A., H. Boubaker, M. Kherallah, A.M. Alimi, A. Ennaji and H. El-Abed, 2009. Arabic handwriting recognition using restored stroke chronology. Proceeding of 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp: 411-415.

El-Abed, H. and V. Margner, 2007. Comparison of different preprocessing and feature extraction methods for off-line recognition of handwritten Arabic words. Proceeding of IEEE 9th International Conference on Document Analysis and Recognition (ICDAR'07), 2: 974-978.

Husain, S.A., A. Sajjad and F. Anwar, 2007. On-line Urdu character recognition system. Proceeding of IAPR Conference on Machine Vision Applications (MVA'07).

Ghods, V. and E. Kabir, 2010. Feature extraction for online Farsi characters. Proceeding of International Conference on Frontiers in Handwriting Recognition (ICFHR'10), pp: 477-482.

Ghods, V., E. Kabir and F. Razzazi, 2013a. Decision fusion of horizontal and vertical trajectories for recognition of online Farsi sub words. Eng. Appl. Artificial Intell., 26: 544-550.

Ghods, V., E. Kabir and F. Razzazi, 2013b. Effect of delayed strokes on the recognition of online Farsi handwriting. Pattern Recogn. Lett. Elsevier Sci. Inc., 34: 486-491.

Hamdani, M., H. El-Abed, M. Kherallah and A.M. Alimi, 2009. Combining multiple HMMs using on-line and off-line features for off-line Arabic handwriting recognition. Proceeding of IEEE 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp: 201-205.

Javed, S.T., S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Moin, 2010. Segmentation free Nastalique Urdu OCR. Proceedings of World Academy of Science, Engineering and Technology, 46: 456-461.

Javed, S.T., 2007. M.A. Thesis, National University, (2007).

Khorsheed, M.S. and H. Al-Omari, 2011. Recognizing cursive Arabic text: Using statistical features and interconnected mono-HMMs. Proceeding of 4th International Congress on Image and Signal Processing (CISP'11), 3: 1540-1543.

Kundu, A., T. Hines, J. Phillips, B.D. Huyck and L.C.V. Guilder, 2007. Arabic handwriting recognition using variable duration HMM. Proceeding of IEEE 9th International Conference on Document Analysis and Recognition (ICDAR'07), 2: 644-648.

Khorsheed M. S. (2007). off-line recognition of omnifont Arabic text using the HMM ToolKit (HTK)," Pattern Recognition Letters, 28(12): 1563–1571.

Kessentini, Y., T. Paquet and A.B. Hamadou, 2010. Off-line handwritten word recognition using multi-stream hidden Markov models. Pattern Recogn. Lett., 31(1): 60-70.

Margner, V., H. El-Abed and M. Pechwitz, 2006. Off-line handwritten Arabic word recognition using HMM-a character based approach without explicit segmentation. Proceeding of Actes du 9` eme Colloque International Francophone sur l'Ecrit et le Document, 2006, pp: 259-264.

Mahmoud, S., 2008. Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models. Signal Pro-cessing, 88(4): 844-857.

Menasri, F., N. Vincent, E. Augustin and M. Cheriet, 2007. Shape-based alphabet for off-line Arabic handwriting recognition. Proceeding of 9th International Conference on Document Analysis and Recognition (ICDAR'07), pp: 969-973.

Mohamad, R.A., L. Likforman-Sulem and C. Mokbel, 2009. Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition. Trans. Pattern Anal. Machine Intell., 31(7): 1165-1177.

Natarajan, P., S. Saleem, R. Prasad, E. MacRostie and K. Subramanian, 2008. Multi-lingual off-line handwriting recognition using hidden Markov models: A script-independent approach. Proceeding of Arabic and Chinese Handwriting Recognition, pp: 231-250.

Natarajan, P., Z. Lu, R. Schwartz, I. Bazzi and J. Makhoul, 2001. Multilingual machine printed OCR. Int. J. Pattern Recogn. Artificial Intell., 15(01): 43-63.

Natarajan, P., D. Belanger, R. Prasad, M. Kamali and K. Subramanian, 2011. Baseline dependent percentile features for off-line Arabic handwriting recognition. Proceeding of International Conference on Document Analysis and Recognition (ICDAR'11), pp: 329-333.

Razzak, M.I., F. Anwar, S.A. Husain, A. Belaid and M. Sher, 2010. HMM and fuzzy logic: A hybrid approach for on-line Urdu script-based languages character recognition. Knowl. Based Syst., 23(8): 914-923.

Razzak, M.I., S.A. Husain, A.A. Mirza and A. Belaıd, 2012. Fuzzy based preprocessing using fusion of on-line and of-fline trait for on-line Urdu script based languages character recognition. Int. J. Innov. Comput. Inform. Control, 8: 1349-4198.

Razzak, M.I., S.A. Hussain, A. Belaıd and M. Sher, 2009a. Multi-font numerals recognition for Urdu script based languages. Int. J. Recent Trends Eng., (IJRTE).

Razzak, M.I., M. Sher and S.A. Hussain, 2010. Locally baseline detection for on-line Arabic script based languages character recognition. Int. J. Phys. Sci., 5(7): 955-959.

Sajedi, H., M. Jamzad, H. Sameti and B. Babaali, 2007. A grouping-based method for on-line Farsi discrete character recognition using hidden Markov model. Proceeding of the 12th International Conference of Computer Society of Iran, pp: 419-426

Satti, D.A. and K. Saleem, 2012. Complexities and implementation challenges in off-line Urdu Nastaliq OCR. Proceeding of Conference on Language and Technology 2012.

Slimane, F., S. Kanoun, J. Hennebert, A.M. Alimi and R. Ingold, 2012. A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution. Pattern Recogn. Lett., 34(2): 209-218.

Slimane, F., R. Ingold, S. Kanoun, A.M. Alimi and J. Hennebert, 2009. A new arabic printed text image database and evaluation protocols. Proceeding of 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp: 946-950.

Naz, S., K. Hayat, M.I. Razzak, M.W. Anwar and H. Akbar, 2013. Arabic script based character segmentation: A review. Proceeding of IEEE World Congress on in Computer and Information Technology (WCCIT), pp: 1-6.

Naz, S., K. Hayat, M.I. Razzak, M.W. Anwar and H. Akbar, 2013. Arabic script based language character recognition: Nasta'liq vs Naskh analysis. Proceeding of IEEE World Congress on in Computer and Information Technology (WCCIT), pp: 1-7.

Xiang, D., H. Liu, X. Chen, Y. Cheng and H. Yao, 2012. Recognition of off-line Arabic handwriting using hidden Markov model toolkit. Proceeding of 11th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES'12), pp: 409-412.