International Journal of Engineering &Technology, 8 (4) (2019) 460-466

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Generating realistic Arabic handwriting dataset

Mahmoud I. Abdalla [1], Mohsen A. Rashwan [2], Mohamed A. Elserafy [3] *

[1]Professor, Electronics and Communication Department, Zagazig University, Zagazig, Egypt
[2] Professor, Electronics and Communication Department, Cairo University, Cairo, Egypt
[3] Engineering, Transit Department, Suez Canal Authority, Ismailia, Egypt
**Corresponding author E-mail: elserafy_eng@yahoo.com*

Abstract

During the previous year's holistic approach showing satisfactory results to solve the problem of Arabic handwriting word recognition instead of word letters segmentation. In this paper, we present an efficient system for generation realistic Arabic handwriting dataset from ASCII input text. We carefully selected sample words list that contains most Arabic letters normal and ligature connection cases. To improve the performance of new letters reproduction we developed our normalization method that adapt its clustering process according to created Arabic letters families. We enhanced Gaussian Mixture Model process to learn letters template by detecting the number and position of Gaussian component by implementing Ramer-Douglas-Peucker algorithm which improve the reproduction of new letters shapes by using Gaussian Mixture Regression. We learn the translation distance between word-part to achieve real handwriting word generation shape. Using combination of LSTM and CTC layer as a recognizer to validate the efficiency of our approach in generating new realistic Arabic handwriting words inherit user handwriting style as shown by the experimental results.

Keywords: *Arabic Handwriting; Normalization; Ligatures; Template Learning; Gaussian Regression*.

1. Introduction

The traditions of Arabic calligraphy were initially inspired by many ways of teaching ancient lines in the Middle East, that is, since the Sumerian era and ancient Egyptian civilization. The Arabic language, which has a number of speakers to 221 million, is considered one of the most difficult languages in the world. The same letter in Arabic have different shapes according to its position in the word. The word in Arabic calligraphy is like the empty container, and the spaces must be balanced within the word itself and within the entire line which result that most of Arabic fonts have ligatures which are a combination of two or more letters [1]. This behavior has led to a great deal of attention has been paid to the search for effective ways of identifying words for large vocabulary using a holistic approach [2], complete words are processed to be recognized bypassing the letter segmentation phase. Part of the challenges facing this approach is large databases for training and testing as well as efficiency in time and space.

The main contribution of handwriting synthesis is to generate text similar to how a human would write the text. Another challenge that Arabic script allows the replacement of certain character sequences by more compact forms called ligatures [3]. Such ligatures lack a systematic analysis despite their importance in Arabic text recognition research.

There are different numbers of techniques used to learn human handwriting movements, can be classified into two main categories Model-free methods, depends heavily on sampling which use auto-encoder and Model-based methods, build template skeleton to generate handwriting by simulating human hand movement depends on few samples of writer style.

In this paper, we present a novel and practical approach for efficient generation of Arabic handwriting letters from ASCII input text. We carefully selected simple word list that contains most Arabic letters normal and ligature connection cases. To improve the performance of new letters reproduction we developed our normalization method that adapt its clustering action according to created Arabic letters families. We enhanced Gaussian Mixture Model (GMM) process to learn letters template by detecting the number and position of Gaussian component by implementing Ramer-Douglas-Peucker algorithm which improve the new letters shapes reproduced by using and Gaussian Mixture Regression (GMR). Also, we learn the translation distance between word-part to achieve real handwriting word generation shape.

The rest of this paper is organized as follows: in Sect. 2, we explore the closely related work of Arabic handwriting synthesis approaches; the proposed approach system is discussed in detail in Sect. 3. Sections 4 present experimental results. Finally, we give conclusion in Section 5.

2. Related work

Research addressing the issue of reproduce human handwriting movements, can be classified into two main categories Model-based and Model-free learning.

Model-free methods learning, depends heavily on large number of sampling which use auto-encoder such as A. Graves [4] long short-term memory recurrent (LSTM) neural network to generate sequences with long-range structure which required a large number of training samples to fit the style or letter classes. Another method of model-free is concatenation technique as Elarian [5] for offline handwritten text, takes character-shape images classified as strictly segmented or extended characters as inputs and concatenates them into synthesized handwriting using the nearest Euclidean-distance neighbor for matching characters that can be concatenated to produce natural-looking words with a WRR of 70.13%. Margner [6] introduced a system by adding an image distortion on the character or word image to simulate the expected real world noise of the intended application. Saabni in [7] for online handwritten generation of words from word-parts is performed based on a predefined layout scheme, which determines the position of the shapes of word-parts with respect to each other and use direct-connection techniques in cursive text lines to synthesize characters. Shatnawi [8] applying congealing technique on Arabic letters to have similar characteristics such as size, position, and rotation, of samples and then use distortion models to synthesize handwritten examples. Most of presented letter-segmented and concatenation technique ignore learning writer handwriting style statistics.
Model-based methods learning, building template skeleton to generate handwriting by simulating human hand movement depends on few samples of writer style. A list of approaches to model letters shape such as A. Almaksour [9] using sigma-lognormal model, to generate handwriting by simulating human hand movement, the model can generate letter samples, applying deformations on sigma-lognormal profiles level allows obtaining modified profiles, which produces unrealistic shapes in many cases. Y. Zheng [10] build letter model represented as a set of points randomly selected from its skeleton, Point matching is used to learn the shape deformation characteristics which used for handwriting synthesis. Dinges [11] use active shape models for generation of Arabic letters by storing most important point information of many shapes of a class in one single model then uses average μ and variance σ of a Gaussian distribution for all affine transformations of ASM build shapes to achieve variations such as slant.
Arabic script is cursive making it viable to support different geometric shapes overlapping and composition. In contrast to Arabic handwriting synthesis, little research has been done concerning overlap between sub-word parts and it is one of the main features that make the Arabic generated words appear natural.

3. Our Approach

We propose an efficient system for generating realistic Arabic handwriting words dataset figure [1]. The proposed system consists of three main components data collection phase, learning phase, and generation phase.

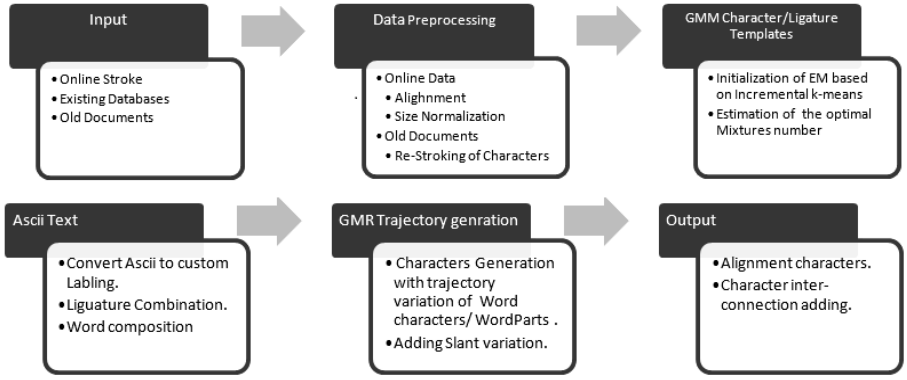

Fig. 1: Proposed Architecture Over View.

3.1. Data collection phase

The main challenge in data collection is selecting the minimum list of words that contains most of Arabic language characteristics such as: all of Arabic letters in different positions and all of mandatory ligatures and famous ligatures.
Arabic language has special characteristics :

1) Writing direction from right to left.
2) The number of basic characters is 28 characters, but the shape of each character increases according to its position in the word isolated - beginning - middle - end.
3) Many ligature occurs where two or more graphemes or letters are joined as a single glyph.
4) Position overlap between sub-word parts.

Arabic ligature style differentiate principally [3] into one mandatory ligature 'lam' character followed by an 'alif' (لا) and typographical ligatures, figure [2] shows that the connection of letters in Arabic handwriting style can lead to implicit contextual ligatures such as two or more letters in the handwritten Arabic language can be combined vertically and represented by different shapes which are considered as new letters classes and we will refer to those letters as compound letters.

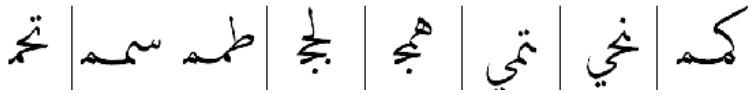

Fig. 2: Typographical Ligature.

One of the characteristics of Arabic handwriting is the overlap between letters within the same word which is famous between most writer the overlap of letters (ر و ز ؤ) with the next letter and also, letters (ج ح خ ع غ) when it comes in the end of the word. So, we use our graphical interface to convert offline word samples list into online word segmented letter samples and calculate the translation distance and angle between overlapped letters figure [3].

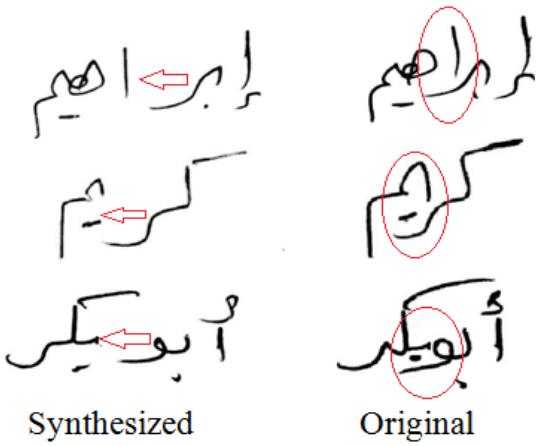

Fig. 3: Shows the Difference between Original and Synthesized Words.

Figure [4] shows sample list of words, which containing most letter joining cases and most ligature using our UI which design carefully to collect samples to extract features of the Arabic handwriting style, trajectories of letters are collected with online techniques since letters features information can be extracted more efficiently from trajectories than images.

Group Name	Letters Description	Example
Isolate Letters	Letters follow them will be in initial ا د ذ ر ز و	بادر – بزور
Mandatory Ligature	Writer have to do it لـ - ا	للكن – ثلاثة
Optional Ligature	letter become before connected vertically هـ – مـ – حـ - جـ - خـ - يـ	بحم – بجح – جميع – مجموع – ضجيج
Special Ligature 1	Repeated of Baa-like letters بـ - تـ - نـ - يـ	بينة – بباب – تثبيت – سبب
Special Ligature 2	Kaff followed by Lam, Haa-like, Meem كلـ - كحـ - كـ	كلمات – كمال – كحروب – كهف
Special Ligature 3	Lam followed by Haa-like, Meem لـ - مـ	لمح – سلمى – مكملة
Overlap Letters	و – ر- ز	الأبواب – وردة

Fig. 4: Example of Sample List.

Nevertheless, we use our UI to easily convert existing dataset and historical offline data to minimize manual effort and allow a simple conversion method. Such as we manually assign the essential control points using Bezier spline curve function to find out the elementary control points given points P0, P1, and P2 shown in figure [5], we can construct a curve P(t) by the following:

$$p(t) = (1-t)^2 p_0 + 2t(1-t)p_1 + t^2 p_2 \quad (1)$$

for $t \in [0,1]$

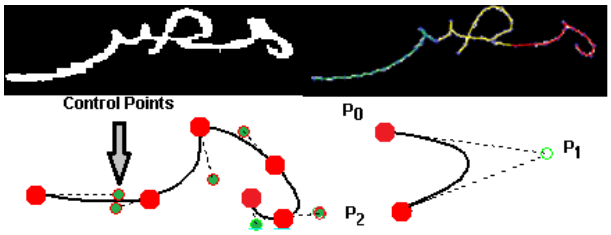

Fig. 5: Shows Bezier Curve Control Points.

The letter trajectory points are placed in a series of Bezier arcs one after another over a character glyph in a order [12] as shown in figure [7]. Each sub-word trajectory points collected as standalone list which will help later in, normalization process, matching main shape trajectory for letters family.

3.2. Data normalization

Dataset normalization is an important step in letters classes learning process, can be even considered as a fundamental building template of each class. As we have studied so many research article, the researchers use standard normalization technique related to dataset, then must of the dataset are not well structured or dataset are unstructured [13].

To be more precise to gather feature information of each class, collected data need to be scaled into common range and due to the nature of Arabic language that some of the letters of the divided into main part and secondary components (sub-word), the secondary component change its related position with the main part every time writer repeat the same letter and by implementation popular scaling normalization methods such as Min-Max, Z-score and Decimal Scaling on letters samples leads to deformation on class template shape of the secondary component beside its relative position to the main part due to using same scale ratio for both parts [14]. To overcome these new challenges of deformation shown in figure [6]. We develop customized normalization method that supports a clustering methodology that enhance Arabic letters classes normalization process, which creates sub-groups of letters according to their main shape trajectory.

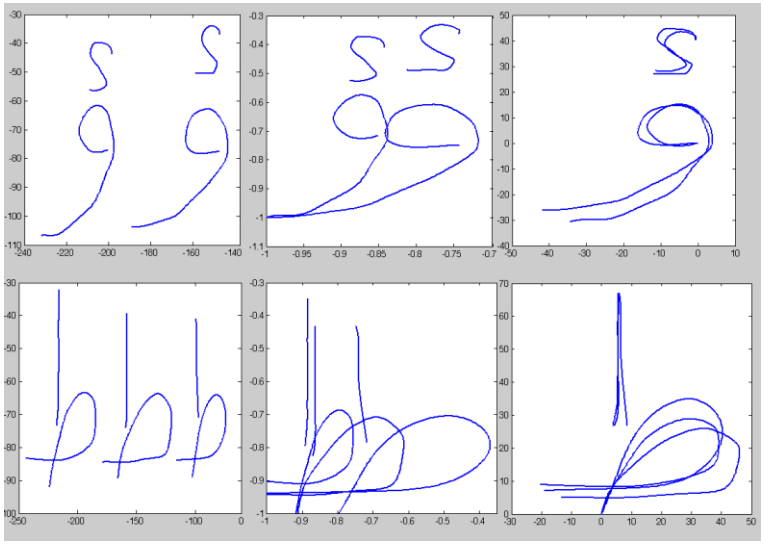

Fig. 6: A) Collected Samples, B) Max-Min Normalization Method, and C) Our Customized Method.

The main contribution of our algorithm is dividing Arabic letters according to similarity of letters main shape trajectory into groups such as one-part set of letters, two-part set of letters, and set of letters similar to the main part.
Our proposed normalization technique having feature of individual letter elements scaling or transformation technique that improves the effectiveness and the performance of normalization algorithm. So, all matched components of all letter groups are aligned then clustered as max-min normalization technique. This technique is a strategy that linearly transform the outputs from one range of values to a new range of values. The rescaling is achieved using the linear transformation
given as:

$$y = \frac{(x - \min(x))}{(\max(x) - \min(x))} \tag{2}$$

Figure [7] gives the outline of our customized normalization model consists of three steps, translation mapping step, which calculates the distance and angle values between main letter part and its secondary components, normalization step, and mean trajectory template of clustered letters elements.

```
Input : letter samples X
Output : normalized trajectory Y template for each letter class
for all letter sample X do
        Assign each letter to its predefined group
        If letter elements > 1
              Compute translation distance between each letter elements
              Align matched letter elements for each letter group
              Scale matched letter elements using max-min normalization
       else
             Align letter for each letter group
             Scale letter using max-min normalization
        end
end

 Calculate Mean trajectory template for each clustered letter elements
```

Fig. 7: Proposed Normalization Model.

3.3. Learning phase

In learning phase, our proposed approach learn local feature of each letter class using Gaussian mixture model (GMM) to build letters templates for each letter class. Nevertheless, To have a realistic handwriting generated words, the statistical variance tolerance between letters need to be learned also, due to the nature of Arabic cursive handwriting making it enable different geometric shapes overlapping between letter elements.

3.3.1. Letter template features extraction

We use Gaussian mixture model (GMM) to build letters templates that inherit writer style features for the calculated mean trajectory model for each letter class, the main contribution is to adapt mean centers, covariance's, and number of components of a GMM of multiple samples for same object. Our proposed technique for Gaussian model goes as follows:

1) Calculating Mean Trajectory Model.
2) Estimating number of model components.
3) Define each cluster by generating a Gaussian model, using EM process to iteratively update the model parameters until convergence.

Mean trajectory model, we cluster similar samples word-part after applying its own scale value. Then compute the means covariance matrices trajectory on the clustered distribution of each normalized class. Implementing of the letters group families figure [8] during the normalization process, that increase the prior probability knowledge of secondary components related to main part cases which is useful to repeat the main shape of letters group which increase writing style features samples of each writer by using small number of collected samples.

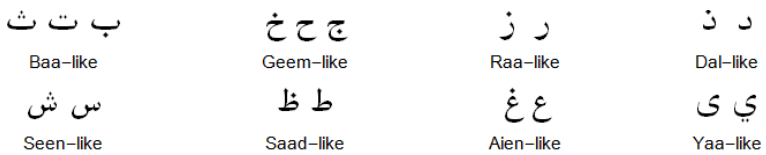

Fig. 8: Proposed Arabic Letters Families.

Estimating number of model components, to estimate the optimal number of mixtures, there have been several [15 -18]. The adapt of mean centers, covariance's affect the feature of the Arabic letters shape. Based on the number and location of the mean center the efficiency of reproduction different models of the letter inherit the writer style handwriting figure [9]. As the number of GMM increases, there will be letter shape over representing, which leads to the fact that all the generated letters derived from generated template are completely identical to the original sample, which is contrary to the main objective technique or by reducing the number of GMM components leading to the reproduction of letters shape dissimilar to writer handwriting style.

We proposed customized technique for estimating optimal number of Gaussian mixtures components based on Arabic letters shapes and writer writing style. Local feature for letter is represented by strongest points that describe the shape of the letter. We use Ramer-Douglas-Peucker (RDP) detection algorithm developed by Urs Ramer in 1972 and proposed by David Douglas and Thomas Peucker in 1973 [19-20]. The Ramer-Douglas-Peucker algorithm is an algorithm to reduce the amount of points in the curve and to represent the original curve with fewer points.

The purpose of the algorithm is to locate the important turning points in the line direction of the calculated mean trajectory model for each letter class. First, letter is divided by the number of syllables per letter elements. Second, RDP algorithm search for the farthest point (Pf) of the letter elements between the start and the end points, If that point is closer than the threshold all points between P1 and Pt are discarded. Otherwise the Pf is included in resulting set. Third, the algorithm repeat the same step recursively with the right and the left parts of the curve (from P1 to Pf and from Pf to Pt). Then merge the results of processing the left and the right parts. Algorithm repeats until all points are handled. Finally, force locating the mean center of Gaussian for each letter elements according to the result of RDP algorithm.

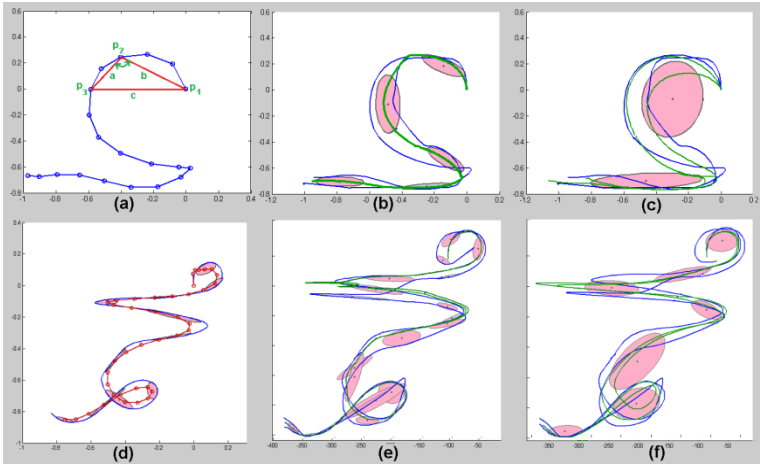

Fig. 9: A-D) Different Letter Curvature, B-E) Over Estimation GMM Component, Green Lines Represent Generated New Letters Looks Identical of Original One, C) Under Estimation of GMM Component, Generated New Letters Looks Different Than of Original Style. F) Good Estimation of GMM Components Using RDP Algorithm.

Finally, learning model for each letter class template is achieved with problem of maximizing the log-likelihood, GMM trained using the Expectation Maximization (EM) algorithm to find maximum likelihood parameters. Maximum Likelihood Estimation of the mixture parameters is performed iteratively using the EM algorithm [21]. EM is a local search algorithm that guarantees finding a locally optimal fit of Gaussians to the data through increasing the likelihood of the training set during optimization constrained to calculated previously means position. A GMM of K Gaussians is defined by the probability density function:

$$p(x_i) = \sum_{k=1}^{K} p(k) \times p(x_i|k), \tag{3}$$

$$p(x_i|k) = N(x_i; \mu_k, \Sigma_k) \tag{4}$$

Where x_i is the datapoint, $p(k)$ is the prior, and $p(x_i|k)$ is the conditional probability density function and μ_k, Σ_k are parameters of the Normal Gaussian distribution components means and covariance respectively.

3.3.2. Translation mapping learning

Translation mapping learning, Arabic script is cursive making it enable different geometric shapes overlapping between letter elements. Letters which are consists of two parts the second part usually being points or Hamzah or a alef and their position change from one letter to another. We learn statistically the writer behaviour style of overlap for letter elements.
We calculate the statistics range values of distance and angle feature of secondary components related to main letter elements X_{ij} for total number L samples of letter class k. Therefore, the average of the distance and angle features for two-part letters is

$$\bar{x}_{ij} = \frac{1}{L} \sum_1^L x_{ij} \tag{5}$$

The relative distance variance feature value between two-part letters can be calculated by Eq. [6] which will be used later in handwriting generation model.

$$s_{ij}^2 = \frac{1}{L-1} \sum_1^L \left(x_{ij} - \bar{x}_{ij}\right)^2 \tag{6}$$

3.3.3. Generating phase

We use our customize procedure to map ASCII codes of new words to special letter code fulfil all letter/ligature cases based on writer handwriting style. The number of letter codes is determined during data collection phase based on handwriting letters/ligatures style appeared from samples.
Our proposed approach using Gaussian Mixture Regression (GMR) in order to generate new trajectory for each letter class [22]. GMR derives regression function from the joint probability density GMM of the model input data that able to generate new letters trajectories shapes inherit the feature of training samples using Gaussian Mixture Regression (GMR) which constrains smoothness.
Finally we generate words from the generated new letters trajectories output of GMR to form word-parts within each word on the same base line. Then determined by calculating the average translation distance between different word-parts.

4. Experimental results

We conducted experiments which show the ability of generating Arabic handwriting new words similar to the writer style handwriting by using a small list of samples. We collected handwritten data by 20 persons each asked to write a list of 300 words.
To evaluate the realistic effectiveness of the GMR generated words, we use our proposed approach to regenerate the original words. Figure [10] illustrates a sample of words that verify the ability of our approach system to generate words from a template model that reinforces the features of the author's handwriting style. Also, the statistical range of the transition distance between the elements of the word-part is taken into account in generation phase.
For further test evaluation, we save the word output as an image samples and ground truth also generated then ran a classification algorithm [23] that would classify the new words generated by our proposed technique, which use Long Short Term Memory (LSTM) Network for sequence learning and Connectionist temporal classification (CTC) that allows the network to make label predictions at any point in the input sequence. The accuracy of the model measured in terms of the overall two error rates, word error rate (WER) and character error rate (CER) were derived. Each was calculated in a similar way where:

$$Error\ Rate\ = \frac{(S + I + D)}{N} \tag{7}$$

Where N is the number of words or characters, S is the number of substitutions (misrecognition of one for another), D is the number of deletions (missed by the recognition system), I is the number of insertions (introduced into the text output by the recognition system).
We use ASCII code to generate 2000 word for training and 400 word for validation including the 300 word writer original words and use 100 new generated word for testing.

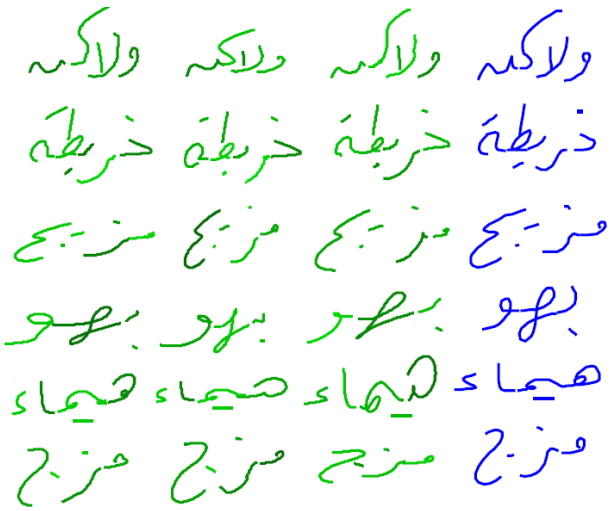

Fig. 10: Shows Samples of Original Handwriting Sample and Generated Samples.

The character error rate for the CTC network was 11.5% with a mean word error rate of 20.4% without using language model.

5. Conclusion

In this paper, we presented a novel approach system for Arabic handwriting generating which extracts the user's style and generate new words inherit writer style. Particularly, our system learned a template model for each letter class wand respect the ligature that varies completely between writers. We learn range distance of writer style overlapped between Arabic word part and enhance the implementation of GMM learned model by using optimal number of components estimation for normalized handwriting samples by using our customize normalization technique. The experimental results demonstrate that our system can produce personal handwriting with realistic visual quality. However, does not capture all aspects of the Arabic handwriting style such as diacritics. our approach offers a valuable approach, which mimics natural handwriting in a better way.

References

[1] Mamoun Sakkal," Arabic Alphabet Chart in Naskh Style", www.sakkal.com.

[2] A. Amin, 2000, "Recognition of Printed Arabic Text Based on Global Features and Decision Tree Learning Techniques", Pattern Recognition, vol. 33, pp. 1309–1323. https://doi.org/10.1016/S0031-3203(99)00114-4.

[3] Yannis H.,1995, "The Traditional Arabic Type-case Extended to the Unicode Set of Glyphs" Electronic Publishing, Vol. 8, pp. 111-123.

[4] A. Graves, "Generating sequences with recurrent neural networks," CoRR, vol. abs/1308.0850, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850

[5] Y. Elarian, Husni Al-Muhtaseb, and LahouariGhouti,2010, "Arabic Handwriting Synthesis", International Workshop on Frontiers in Arabic Handwriting Recognition, Istanbul.

[6] Margner V, Pechwitz M (2001) Synthetic Data for Arabic OCR System Development. In: Sixth International Conference on Document Analysis and Recognition (ICDAR'01), IEEE: 1159-1163.

[7] R.M. Saabni, J.A. El-Sana,2013, "Comprehensive synthetic Arabic database for on/offline script recognition research," Int. J. Doc. Anal. Recognit. (IJDAR) 16 (3) pp. 285–294. https://doi.org/10.1007/s10032-012-0189-5.

[8] Shatnawi M. and Abdallah S.,2015,"Improving Handwritten Arabic Character Recognition by Modeling Human Handwriting Distortions," ACM Transactions on Asian and Low-Resources Information Processing. https://doi.org/10.1145/2764456.

[9] A. Almaksour, E. Anquetil, R. Plamondon, and C. O'Reilly, Synthetic handwritten gesture generation using sigma-lognormal model for evolving handwriting classifiers, in: Proceedings of the 15th Biennial Conference of the International Graphonomics Society, 2011, pp.98–101.

[10] Y. Zheng and D. Doermann, "Handwriting matching and its application to handwriting synthesis," in Proceedings of the Eight International Conference on Document Analysis and Recognition (ICDAR), 2005, pp. 861–865.

[11] Dinges, L.; Al-Hamadi, A.; Elzobi, M.; El etriby, S.; Ghoneim, A. ASM based Synthesis of Handwritten Arabic Text Pages. Sci. World J. 2015, 2015, 323575. https://doi.org/10.1155/2015/323575.

[12] D. Salomon, "Curves and Surfaces for Computer Graphics", Ch.1, pp.7-14, Springer, 2006.

[13] Mustaffa and Yusof. A Comparison of Normalization Techiques in Predicting Dengue Outbreak. International Cinference on Business and Economics Resaerch, vol.1(2011) © (2011) LACSIT Press, Kuala Lumpur, Malaysia.

[14] Patel and Mehta. Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011, ISSN (Online): 1694-0814.

[15] G.Schwarz, "Estimating the Dimension of a Model," Annals of Statistics, vol. 6, 1978, pp. 461-464. https://doi.org/10.1214/aos/1176344136.

[16] C. Biernacki, G.Celeux and G. Govarert, "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," Technical Report 3,521, Inria, 1998.

[17] A.Likas, N.Vlassis, and J.Verbeek, "The Global k-means clustering algorithm," Pattern Recognition 36, 2003, pp. 451-461. [12] J.Verbeek, N.Vlassis, and B.Krose, "Efficient Greedy Learning of Gaussian Mixture," Neural Computation 15, 2003, pp. 469-485. https://doi.org/10.1016/S0031-3203(02)00060-2.

[18] Y Lee, KY Lee, J Lee.,2006, "The estimating optimal number of Gaussian mixtures based on incremental k-means for speaker identification", International Journal of Information Technology 12 (7), pp13-21.

[19] U. Ramer, An iterative procedure for the polygonal approximation of plane curves, Computer Graphics and Image Processing 1(3) (1972) 244-256. https://doi.org/10.1016/S0146-664X(72)80017-0.

[20] D.H. Douglas, T.K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, Cartographical: The International Journal for Geographic Information and Geovisualization 10(1973) 112-122. https://doi.org/10.3138/FM57-6770-U75U-7727.

[21] A. Dempster and N. Rubin, "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal Statistical Society, vol. 39(1), pp. 1–38, 1977 https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[22] D. Cohn, Z. Ghahramani, and M. Jordan, Active learning with statistical models. Articial Intelligence Research, vol. 4, pp. 129145, 1996. https://doi.org/10.1613/jair.295.

[23] Alex Graves and Jürgen S. 2009, "Offline handwriting recognition with multidimensional recurrent neural networks". In Advances in Neural Information Processing Systems 21, pp 545-552.