Feature Level Fusion for Online Arabic Script Based Languages Character Recognition

Muhammad Imran Razzak
Department of Computer Science and Engineering
Air University, Islamabad
Pakistan
imranrazak@hotmail.com

Mohammed Talal Simsim
College of Engineering at Al-Lith
Umm Al-Qura University, Makkah
Kingdom of Saudi Arabia
mtsimsim@uqu.edu.sa

***Abstract*—Arabic script character recognition remains a challenging task due to its cursive nature. The aim of feature extraction from the input strokes reduces the input pattern to avoid complexities while maintaining the high accuracy. Feature extraction in pattern recognition problems involves the extraction of unique and salient patterns from the preprocessed data in order to enhance the discriminatory power and reduce the data for classification and it is crucial for the success of recognition system. We proposed feature level fusion for Arabic character recognition. We extracted structural and directional features for handwritten stroke and fused these feature to form more discriminant feature matrix. The feature level fusion for handwritten character recognition provides considerable gain in accuracy. The fusion of feature is also suitable for other handwritten tasks i.e. personality determination, writer identification where the more variations are required.**

Keywords*—Feature Extraction, Arabic Character Recognition, Feature Fusion.

I. Introduction

Character recognition is an important offshoot of pattern recognition problems. It imitates a human's ability to read, using a machine. It has been a field of intensive, if exotic, research since the early days of the computer. This task becomes more complex and demanding in case of handwritten and cursive text. Arabic script-based languages are used by almost a quarter of the world's population [1]. Character recognition has been classified into two categories based on the mode of input: online character recognition and offline character recognition. Offline handwritten recognition does not require immediate interaction with the user while online handwritten recognition has complete interaction with the user. The root of online handwriting recognition is real time data. For online character recognition, commonly input devices are digitizing tablets or digital pen, where the written data is digitized and translated into a series of coordinates.

Chinese and Roman script have tremendously attracted interests of character recognition researchers both for online and offline input while in contrast, the research efforts for Arabic script based languages are very less. This may be due to the complexities of this script over Roman and Chinese script. For example, Arabic script has multiple shapes of one character that depends upon the position of the character in ligature. Arabic script based languages are mostly written in two writing style Naskh and Nastaliq. Urdu character set is the super set of all Arabic script based languages [1]. Nastaliq is a special calligraphic way of writing and is mostly used especially for handwriting.

The direct recognition of handwritten stroke is almost impossible due to high variability of handwritten strokes. The feature selection phase also called dimensionality reduction. The aim of feature extraction from the input strokes is reducing the input pattern to avoid complexities while keeping the high accuracy and the extraction of those distinct patterns that uniquely define the strokes and most important for classification while the task of human expert is to select those features that allow effective and efficient recognition. Feature extraction crucial for the success of recognition system. We proposed feature level fusion for Arabic character recognition. We extracted structural and directional features for handwritten stroke and fused these feature to form more discriminant feature matrix. The information of each pattern is recorded as similar to human visual concept. The feature level fusion for handwritten character recognition provides considerable gain in accuracy. The fusion of feature is also suitable for other handwritten tasks i.e. personality determination, writer identification where the more variations are required. Section II introduces the literature work, in section III Feature Extraction and fusion of feature is performed. Results are discussed in section IV.

II. Related Work

Within the context of online handwritten character recognition, studies dealing with Arabic characters are scarce [2]. The main aim of feature extraction phase is to extract those unique patterns which are most pertinent for classification and uniquely describe the ligatures. The feature extraction phase must be robust enough so that the extracted features are small in numbers produce less error during extraction.

J. Sternby et al. extracted relative features i.e. relative horizontal position, relative vertical position, with the mean value of segments and corresponding mean horizontal value and each segment is identified by feature angle, arc type, connection angle, length ratio and relative position [3]. A.

SEEK DIGITAL LIBRARY

Borji et al. extracted simple cells and grow these cells by connecting the associated cells with each other to form the features [4] according to biological visual perception. Malaviay and L. Peters presented a layered approach by extracting features from level 0 (strokes elements) towards unique identified patterns. From the strokes elements, geometrical features are computed by using the fuzzy membership function, and these small geometrical features are combined to form global features by using fuzzy primitives [5]. In order to obtain 24 dimensional feature vectors J. Schenk extracted three dimensional feature vectors from both offline and online domain [6]. The online features are pen pressure, velocity, (x ,y) coordinates, difference of angles, angle between lines etc. and for offline features, the strokes is sub sampled into 3x3 along pen trajectory and ascenders and descenders [6].

El-Anwar et al. proposed three types freeman chain codes, long strokes, short strokes and eight pen-up for ensample movement with additional pen down information of succeeding stroke and current stroke [6]. Malik and Khan used three directional features with other features like slope, writing direction, size start and ending coordinates [8]. Hussain et al. extracted 20 unique shape defining features from primary stroke. [9]. Benouareth et al. used uniform and non-uniform segmentation scheme and extracted both statistical and structural features for offline Arabic character recognition [10]. Beta-elliptical representation is also used for features extraction from the handwritten strokes [11-13] and in the second phase Kherallah et al. transformed the beta-elliptical model trajectory by visual codes based on psychological and cognitive domain and the curvilinear velocity is computed using a second-order derivative. Husam et al. presented efficient neural based segmentation for Arabic handwritten word recognition and the modified direction feature extraction technique combines the local feature vector and global structural information. First the contours are extracted and then directions of line segments comprising the characters are detected and the foreground pixels are replaced with the direction values [14]

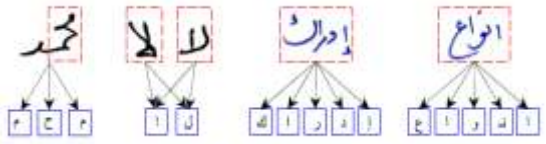


Fig. 1. Arabic ligature and their constituent character [7]

Sternby et al. extracted relative horizontal and relative vertical position by signifying the mean vertical value of segment and each segment is represented by specified feature angle, arc type, length ratio, connection angle and the relative position shown in figure 2 [3].

Borgi used the Gabor and DoG filters as in the feature extraction phase to implement biological inspired system [4]. Izadi et al built shape descriptors and define relative context as in the pair wise distance and angle and rational property spanned by relative context making it possible to include different levels of detail [15]. Amor and Amara extracted features using Hough transformation which correspond to major line features [16]. Aly et al. used a technique that normalized the text in bounding box by estimating the partial height of character using the normalized size and height of the characters are defined [17]. Hanmandlu et al. presented box method for feature extraction for offline unconstrained character recognition and 24 pairs are extracted from handwritten text and all the features are organized in sequential order [18].

Khedher and Al-Talib combined statistical and structural method for feature extraction and 12 features are extracted from both main and secondary parts of the characters. For secondary stroke limited features such as height, width, height to width ratio are extracted [19]. Baghshah et al. extracted directional features and relative vertical and horizontal motion features from online handwritten for Persian character recognition. [20]. Husam et al. presented segmentation based approach combined with neural network for handwritten Arabic recognition [21]. The words are segment into uniform and non-uniform segmentation. The segmentation points are validated using neural network by fusing the confidence value and several directional features (manifold direction, normalized direction value etc.) are extracted for successful segmentation.

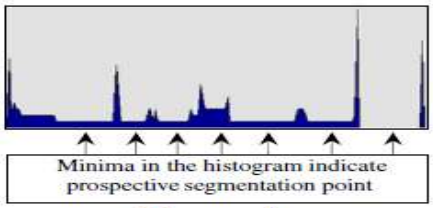


Fig. 2. Segmentation point estimation based on histogram [3]

The above review presents the feature extraction phase. The literature shows that structural and directional features are important for handwritten classification. Extraction of both structural and directional features increases the dimension of the feature matrix. To reduce the size of the feature matrix and makes it more robust we present the feature level fusion of both direction and structural features based on the time and location information of each feather for online Arabic script character recognition.

III. Feature Extraction

A successful character recognition methodology depends on the choice of features used by the pattern classifier. Feature extraction in pattern recognition problems involves the derivation of salient features from the preprocessed data in order to reduce the amount of data used by the classifier for classification and simultaneously provide the enhanced discriminatory power. The feature extraction phase must be robust enough so that extracted features are small in number, produce less error during extraction and uniquely describe the shape of strokes. Generally, handwritten strokes are oriented lines, loops, and curves. These orientations play an important role in the classification of strokes. The directional and structural features i.e. loop, cusp, etc. play important role in the classification and previous work shows that improved

SEEK DIGITAL LIBRARY

results have been obtained by combining structural and directional features. Structural features are the shape defining features and these are based on the instinctive aspects of writing and include loops, cusp, endpoints, starts points etc. The and where information of each pattern is recorded as similar to human visual concept. Sometimes, further preprocessing is required on this feature matrix to remove the unnecessary and noisy features by using language rules. Time variant structural feature are extracted from both online and offline stroke elements to identify on-line handwritten ligatures. The additional time information is utilized to extract features from the stroke elements in the order of occurrence. This arrangement helps us in combining the diacritical marks through estimation. Finally, post processing is applied on extracted features matrix to combine the features and remove the unnecessary or noisy features based on language rules. Figure 3 briefly describe the directional feature extraction process.

Handwritten strokes are oriented lines, curves, or poly-lines which play an important role in differentiating between various characters. For a long time, orientation or direction has been taken into account in handwritten character recognition. In early stages, character recognition using directional features was called directional pattern recognition [Fujisawa and Lui, 2003]. Chain code is the important way to extract the directional features. We have used chain codes and length for directional feature extraction based on fuzzy logic and context knowledge of the ligature itself, instead of length of current feature as shown in figure 4. The relative fuzzy decision of directional features is performed in two steps. The first stage extracts small patterns and the second stage combines these small patterns to form large patterns. The decision of large patterns is performed in level-4 based on the associated patterns detected in the ligature using the fuzzy language rules.

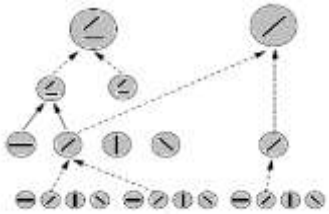
As humans have high context knowledge and they can better differentiate the different sizes based on their context knowledge, similarly we have tried to model the context knowledge of ligature itself. Although it is very limited yet it helps to differentiate and extract more discriminating features.

A. Directional Feature Extraction

Handwritten strokes are oriented lines, curves, or poly-lines which play an important role in differentiating between various characters. For a long time, orientation or direction has been taken into account in handwritten character recognition. In early stages, character recognition using directional features was called directional pattern recognition [Fujisawa and Lui, 2003]. Chain code is the important way to extract the directional features. We have used chain codes and length for directional feature extraction based on fuzzy logic and context knowledge of the ligature itself, instead of length of current feature as shown in figure 4. The relative fuzzy decision of directional features is performed in two steps. The first stage extracts small patterns and the second stage combines these small patterns to form large patterns. The decision of large patterns is performed in level-4 based on the associated patterns detected in the ligature using the fuzzy language rules.

As humans have high context knowledge and they can better differentiate the different sizes based on their context knowledge, similarly we have tried to model the context knowledge of ligature itself. Although it is very limited yet it helps to differentiate and extract more discriminating features.

- Start_Small_Directions. This feature depends upon the small movement at start of the ligature either left, right, top, down or diagonal direction e.g. ا،ب.
- Start_Vertical_Down. This feature is selected when the ligature was a straight long vertical downward in the beginning. e.g.ط،ل،١.
- Start_Vertical_Up. This feature is selected when the ligature was a straight vertical upward in the beginning. As there is no word which starts from upward but this feature is used to differentiate numerals like ١ and ا having same shapes.
- Ending _Vertical_Long _Up: This feature is selected when the ligature was a straight long vertical upward in the end. For e.g.یا، کا طا.
- Ending_Vertical_Long_down: This feature is selected when the ligature was a straight long vertical downward in the end. e.g.ہم م .
- Long_Horizontal_Left: This feature is selected if during writing the ligaturethe pen movement is very long and from right to left horizontally e.g. inب، ف.
- Long_Diagonal_Left: This feature is selected if during writing the ligature the pen movement is long and from left to right diagonally like in لا.
- Long_ Diagonal _Right: If during writing the ligature, the pen movement is long and from left to right horizontally then the horizontal left to right is selected e.g. in ے، سے



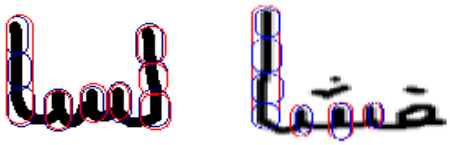


Fig. 3. Directional features

B. Structural Feature Extraction

- Some shapes have semi-circle. For such characters, feature called the semicircle are extracted when left to right semicircle present from right to left like ق،ن.
- Similarly the ligatures containing the semicircle from

SEEK DIGITAL LIBRARY

left to right i.e. ع ج.

- The direction of writing of curves varies from right to left and from left to right. Therefore, curve right to left is selected for characters having shapes like ر ن .
- This feature is selected if the curve direction is from left to right like ع ج .

$$\alpha = \Theta_{(i+1,S)} - \Theta_{(i-1,S)}$$

$$\Theta_{i,S} = Arc\tan\left[\frac{(y_i - y_{(i-k)})}{(x_i - x_{(i-k)})}\right]$$

- Cusps are the sharp turning point in a stroke. This feature is selected for the ligature which contains the up side cusps such as those present in س, سر.

$$\alpha = \Theta_{(i+1,T)} - \Theta_{(i-1,T)}$$

$$\beta = \Theta_{(i+1,T)} - \Theta_{(i,T)}$$

$$\beta - \alpha < 8$$

- This feature is selected for the ligature which contains the downward cusps such as those present in سہہ،بہب .Similarly using the above value of α, β.

$$\alpha - \beta < 8$$

- Whenever an intersection is encountered in a primary stroke this feature is selected i.e. ط, فل .
- This feature is selected if there is ray shape at the end of the stroke. If any ligature is a combination of ray or dal then this feature is also selected i.e. ر،بد،د.
- In order to differentiate the loop the clockwise written loop is selected i.e. ف,ق .
- In order to differentiate fay (ف), qaf (ق) and meem (م), this feature is selected for Meem. The writing direction of the loop in meem is anti-clockwise. i.e. بم،جم تم.
- To differentiate the loop in Sad (ص) from the other loop, this feature was identified and selected for Sad. As the sad loop is egg shaped so it is identified to separate the Sad from other loops like ص بص. Some time it is very difficult to differentiate between the loop-swad, loop-down, circular hey (ھ) and loop-up. If loop selection problem occurs then final decision is based on the fusion of directional and structural features.
- In order to differentiate the loop in fee, Qaf, Meem and hey , this feature is identified when isolated loop occur i.e. Hey "ھ" "ہ".

Table 1 Recognition result based on feature level fusion.

Approach	Input Type	Data Set A (Nastaliq)	Data Set B (Naskh)
Hybrid Approach (HMM + Fuzzy)	Primary Stroke	91.4	84.8
Hybrid Approach (HMM + Fuzzy)	Ligature with mapped diacritical marks	90.6	84.1

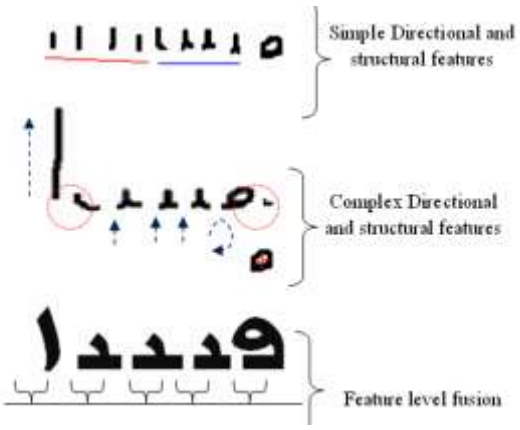


Fig. 4. Feature level fusion of directional and structural features

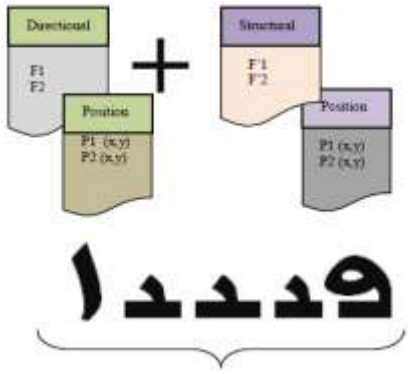


Fig. 5. Fusion based on location information of each feature

- This feature is selected if some characters combine with ray (ر) like برجر .
- This feature is very difficult to extract because there ~~are~~ is very small variation involved. It is selected if some characters combines with choti ye like جیبی.
- This feature is used to resolve the problem in differentiating between jeem (ج) and ayen (ع) i.e. عب، ع. Its selection depends upon the previous features and extracted having using the following equation.

$$\beta = \Theta_{(i+1,S)} - \Theta_{(i,S)}$$

$$150 > \alpha < 180 \;\; \&\& \;\; 200 > \beta < 250$$

- This feature is selected on the presence of tuan in any ligature like ط، سط بطا. Tuan is selected when loop exists after long up and down.
- When character combines with hey this feature is selected. i.e. جہ بہ .

SEEK DIGITAL LIBRARY

C. Fusion

Fusion of structural and directional features is performed based on relative position information of each feature. Z is the feature and x,y are the feature coordinates. Where and what information of each salient pattern is extracted and used for feature fusion. The features are combined to form new features based on the language rules. For example the directional feature Start_Vertical_Down is combined with structural feature Loop to form new feature for ligature ﻟ. This new feature is more discriminant as compared to the previous two features because it utilizes the location information as well in the construction of new feature matrix.

$$F_S = \{ f_1, f_2, \ldots\ldots f_n \}$$

$$F_D = \{ f_1^{'}, f_2^{'}, \ldots\ldots f_m^{'} \}$$

Where $f_i = (Z_i, x_a, y_b)$

$$f_j^{'} = (Z_j, x_a, y_b)$$

$$F_F = \{ f_1^F, f_2^F, \ldots\ldots f_k^F \}$$

The fused feature matrix is given as

$$F_F = \{ f_1^F, f_2^F, \ldots\ldots f_k^F \}$$

IV. Conclusion

As successful character recognition methodology depends upon the particular choice of features used by the classifier. Thus it is the most critical part in any recognition problem. Due to the complexity and variation, direct recognition of handwritten stroke is almost impossible. The literature shows that directional and structural features i.e. loop, cusp etc. played important role in the classification. Fuzzy rules have been used to extract the unique and meaningful directional and structural features and shape defining patterns i.e. loops, cusp, endpoints, starts points, etc. Further post processing is also applied on extracted features to remove the unnecessary and noisy features by modeling the language rules and fusion of directional and structural features. Where and what information of each salient pattern is extracted and used for feature fusion. The features are two close points are combined to form new features. For testing purpose, we used the system [22] and it gains a considerable accuracy by using the feature level fusion. The approach was based on both directional and structural features. We fused both directional and structural features and it provides accuracy.

References

[1] Razzak M.I., S.A.Hussain,A.Belaid, M.Sher (2010) "Multifont numeral recognition for Urdu script based languages " International journal of research trend in engineering.

[2] Inam Shamsher et.al, OCR For Printed Urdu Script Using Feed Forward Neural Network, Proceedings of World Academy of Science, Engineering and Technology. Vol 23, Aug 2007 ISSN1307-6884

[3] Sternby J., Morwing J., Andersson J., Friberg C., (2009) "On-line Arabic handwriting recognition with templates" Pattern Recognition 42 3278 – 3286

[4] Borji A., Hamdi M, F Mahmoudi, (2008), "Robust Handwritten Character Recognition with Features Inspired by Visual Ventral Stream" Neural Processing Letters Vol 28 pp 97-111.

[5] Malaviya A., L. Petrs,(1999), "Multilayered Handwritten Recognition Appraoch", Fuzzy Sets and System Vol 104, pp 219-227

[6] Schenk J., S. Schwarzler, G. Rigoll, (2008), "PCA in Online Handwritten Recognition of Whiteboard Notes: A Novel VQ Designe for Use with Discrete HMM's" in: Proc. of Int. Conference on Frontiers in Handwriting Recognition S. 544 – 549

[7] Al-Hamad H.A., Zitar R. A. (2010) "Development of an efficient neural-based segmentation technique for Arabic handwriting recognition" Pattern Recognition 43 2773–2798

[8] Malik, S. Khan, S.A., "Urdu Online Handwriting Recognition", Emerging Technologies, 2005. Proceedings of the IEEE Symposium on Volume, Issue, 17-18 Sept. 2005 Page(s): 27 – 31

[9] S.A.Hussain , Anwar F., Asma. "Online Urdu Character Recognition System." MVA2007 IAPR Conference on Machine Vision Applications.

[10] Benoureth A, Ennaji A, Sellam M. (2008)"Semi-Continuous HMMs with Explicit State Duration Applied to Arabic Handwritten Word Recognition , Pattern Recognition Letter pp 1742-1752.

[11] Kherallah M, Bouri F., Alimi A.M.,(2009) "On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm" Engineering Applications of Artificial Intelligence 22 pp. 153–170

[12] Kherallah M., Elbaati A, Abed H.E., Alimi A.M., "The On/Off (LMCA) Dual Arabic Handwriting Database" International Conference on Frontiers in Handwriting Recognition, 2008.

[13] Alama S.' adeed (2008). , Recognition of Offline Handwritten Arabic Word Using Hidden Markov Model Approach" , 16th International Conference on Pattern Recognition (ICPR'02) - Volume 3.

[14] Husam A.A, Zitar R.A.(2010), Development of an efficient neural-based segmentation technique for Arabic handwriting recognition, Pattern Recognition 43 pp. 2773–2798

[15] Izadi S., Sue C.Y. (2008), "Online Writer-independent Character Recognition Using a Novel Relational Context Representation" 2008 Seventh International Conference on Machine Learning and Applications

[16] Amor N.B., Amara N.E.B (2006), "Multifont Arabic Characters Recognition Using HoughTransform and HMM/ANN Classification", Journal of Multimedia, Vol. 1, NO. 2, May 2006

[17] Aly W., Uchida S, Suzuki M (2007), "Identifying Subscripts and Superscripts in Mathematical Documents", Mathematics in Computer Science

[18] Hanmandlu M., Murali Mohan K.R., Chakraborty S, Goyal S,Choudhury D.R. (2003) "Unconstrained handwritten character recognition based on fuzzy logic" Pattern Recognition 36 pp. 603 – 623

[19] Khedher M.Z, Al-Talib G. (2007), "A fuzzy expert system for recognition of handwritten Arabic sub-words" 2007 IEEE

[20] Baghshah M.S., Baghshah S.B., Kasaei S. (2006) "A novel fuzzy classifier using fuzzy LVQ to recognize online Persian handwriting", Information and Communication Technologies, 2006. ICTTA '06

[21] Husam A.A, Zitar R.A.(2010), Development of an efficient neural-based segmentation technique for Arabic handwriting recognition, Pattern Recognition 43 pp. 2773–2798

[22] Razzak M.I., Anwar F, Husain S.A., Belaid A, Sher M, (2010), HMM and fuzzy logic: A hybrid approach for online Urdu script-based languages' character recognition, Knowledge-Based Systems, 2010

SEEK DIGITAL LIBRARY