International Journal of the Physical Sciences Vol. 5(7), pp. 955-959, July 2010
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2010 Academic Journals

Full Length Research Paper

Locally baseline detection for online Arabic script based languages character recognition

Muhammad Imran Razzak[1]*, Muhammad Sher[1] and S. A. Hussain[2]

[1]International Islamic University, Pakistan.
[2]Air University, Pakistan.

Accepted 15 June, 2010

Baseline detection is one of the most important step in character recognition and has direct influence on recognition result. Due to the complexity of the Urdu scripts based languages, handwritten character recognition is a very difficult task as compared to other languages. Baseline detection is one of the main issue and basic step of mostly preprocessing operations that is, normalization, skewness, secondary strokes segmentation and also in feature extraction. This paper presents a novel method of baseline detection for cursive handwritten Urdu script. The proposed approach is divided into three steps: diacritical marks segmentation, primary baseline estimation and local baseline estimation. The local baseline extraction is estimated using the features extracted from ending shape of the words. Due to structural difference between Nasta'liq and Naskh style, different rules are formed for baseline estimation.

Key words: Baseline, Arabic, Nasta'liq, Naskh, preprocessing, character recognition, skewness.

INTRODUCTION

Character recognition has been an on-going research from early days of computer and remains a challenging issue in the field of pattern recognition due to complexities involved specially in handwritten character recognition. With respect to mode of input character recognition is classified in to two main classes offline and online. In offline the input is spatial coordinates in the form of image where as in online additional timing information is available with stroke elements. This additional timing information makes online character recognition a little easy as compared to offline character recognition. From the last few decades online character recognition is getting more popularity as compared to offline character recognition due to increasing popularity of hand-held devices and natural way of input to the machines.

Urdu scripts is almost used by more than one-fourth population of the world in the form of many languages that is, Arabic, Persian, Urdu, Punjabi, Pashto etc. Urdu consists of 58 alphabets and the ghost shapes of Urdu alphabets also exist in other languages. Moreover Urdu is mostly written in Nasta'liq style which is more complex than any other style followed by other Arabic script based languages that is, Nasta'liq contains 32 shapes for 2nd character "ٹ" depending upon the attached character whereas Arabic consist of only four shapes for each character. In other words Urdu is more complex due to diacritical marks in its family. Urdu script based languages are written in cursive style from right to left and are very rich in diacritical marks and it is also the context sensitive language and written in the form of ligatures which comprises single or many different characters to form one ligature. Most characters have different shapes depending on their position and its adjoining character in the word that is, middle, end, start, isolated. Moreover the complexities are increased because the characters overlap each other.

The character recognition for Urdu script based languages is much more complicated than other language like English and Chinese due to complexities of this script that is, context sensitive shape, cursiveness, overlapping, baseline, ligature and space between the ligature. While within the family of Arabic character recognition, studies dealing with Arabic script based languages characters are very less especially for Urdu script.

Preprocessing is one of the most important phase of Urdu character recognition and directly influence on recognition result. Baseline is the virtual line on which

*Corresponding author. E-mail: imranrazak@hotimail.com.

semi cursive or cursive text are aligned/joined and is an ideal parameter to simplify the handwritten text (Boubaker et al., 2009). Baseline detection is one of the most important step in preprocessing of character recognition and has direct influence on efficiency, reliability and complexity of the recognition system. In other words baseline has direct influence on recognition result. Whereas due to the complexity of the Urdu scripts, baseline estimation for text written in Nasta'liq script is a very difficult task. Baseline detection is one of the main issue and basis of mostly preprocessing operations that is, normalization, skewness, secondary strokes segmentation and also in feature extraction. As Urdu script based languages are written in the form of ligatures, thus for online input a ligature may consist of more than one strokes. Baseline is an important parameter to estimate and recombine the segmented ligatures to form one ligature. The main use of baseline is skew correction. During handwriting, user may not write on horizontal straight line thus, orientation is required before further processing. This paper presents a locally method of baseline detection for cursive handwritten Urdu script.

RELATED WORK

Baseline is the virtual line on which semi cursive or cursive text are aligned/ joined. Generally baseline is kept in mind during both writing and reading. Baseline detection is not only used for automatic character recognition but it is also necessary for human reading. Without base-line detection it is very difficult and big issue to read the text even for human and error rate increase up to 10% while the context sensitive interpretation is involved. Whereas in automatic classification no context based interpretation is involved, thus baseline detection is the necessary part of better classification especially for Arabic script based languages. The detection of diacritical marks are not easy task without baseline. Several baseline detection methods based on horizontal projection have been proposed in literatures but they are for large text lines.

The horizontal projection based approach counts the elements on horizontal line and assumes that maximum number of elements on horizontal line is the baseline. Although it is robust and very easy to implement but it needs long straight line of text but in the case of handwritten text especially for online handwritten the length of line may be very short. Thus the histogram projection mostly failed in estimating the correct baseline for isolated handwritten text and ligatures having greater number of ascender and descender.

Boubaker et al presented a novel method for both online and offline Arabic handwritten text (Boubaker et al., 2009). S.S. Maddouri and H.E. Abed compared the six methods for Arabic character recognition on IFN/ENIT database. Projection based method fails to estimate the baseline for short word length and words having more diacritical marks, ascender, descender. Min-Max and PAWs are presented based on the projection based method. Min-Max contour method used critic point from the word contour and two baselines upper and lower are extracted from mean of maxima and minima respectively. The combination of Min-Max and some structural primitives that is, loops and diacritical marks are used for baseline estimation. These additional primitives are used to differentiate contours from the others.

Faisal et al modified the RAST algorithms by introducing two desender line d1 and d2 for Urdu images. Farooq et al presented linear regression on local minima of word for baseline detection (Farooq et al., 2005; Benoureth et al., 2008). R.A. Mohammad et al presented a vertical projection algorithm obtained by summing the value along x-axis and detected two baselines. The lower baseline is identified by maximum projection profile. The upper baseline is estimated by scanning the image from top to bottom (Mohamad et al., 2009). Razzak et al. (2010) extracted baseline by computing the minimum enclosing rectangle and drawing vertical projection (Razzak et al., 2010). Alkhateeb presented knowledge based baseline estimation by using the location information for baseline estimation. The algorithm is improved by estimating the baseline at bottom half of image because of baseline existence at bottom of word [6]. The vertical projection is inefficient for small length text. Benouareth et al used projection after transforming image into Hough parameter for baseline estimation. Pechwitz and Maergner used linear piecewise curves using projections for baseline estimation (Pechwitz and Maergner, 2003).

PROPOSED BASELINE ESTIMATION

Baseline is the virtual line on which characters are combined to form the ligatures and it is the necessary requirement for both readers and writers. Urdu script based languages are written in many styles but mostly nasta'liq and naskh is followed. Urdu, Punjabi etc. is written in Nasta'liq whereas Arabic, Perisan etc are written in naskh style. We proposed a novel technique based on some primitives extracted from the ghost character.

Nasta'liq and Naskh styles are mostly followed by Urdu script based languages. Figure 1 shows the modeling of Urdu on baseline and two descender lines. Different character appears at different descender line. Due to the complexity of Nasta'liq over Naskh, one character may appear at different descender line depending upon the associated characters whereas in Naskh style last character appears on the one baseline and does not depend upon its connected character shown in Figure 2. Thus baseline estimation for nasta'liq written text is more complex than naskh style. Without pre knowledge of word structure it is very difficult to estimate the baseline. The

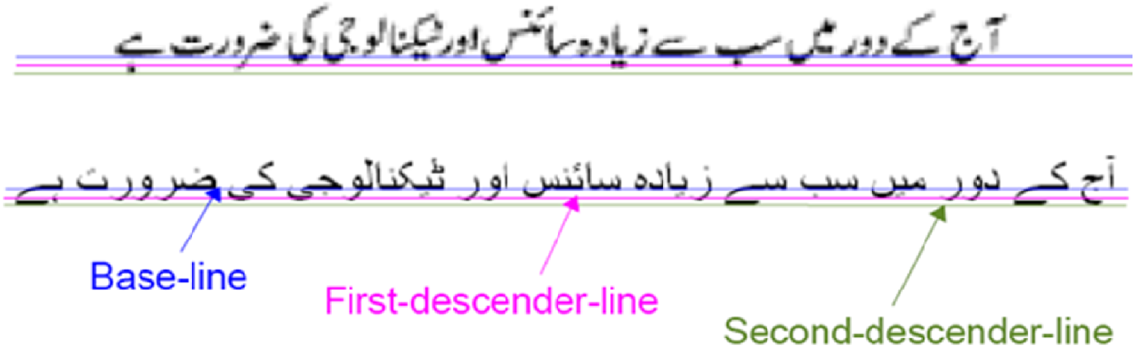
آج کے دور میں سب سے زیادہ سائنس اور ٹیکنالوجی کی ضرورت ہے
آج کے دور میں سب سے زیادہ سائنس اور ٹیکنالوجی کی ضرورت ہے
Base-line
First-descender-line
Second-descender-line

Figure 1. Baseline and descender lines for Nasta'liq and Naskh font for Urdu (Faisal et al., 2004).

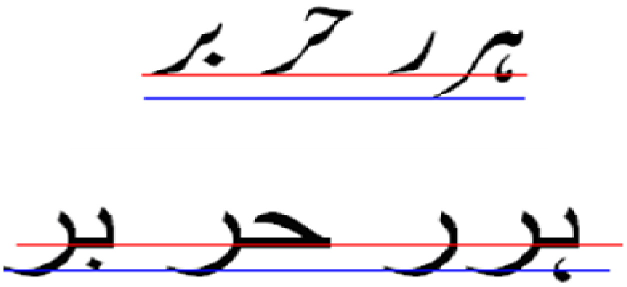

Figure 2. Baseline (Red) for Nasta'liq and Naskh Style and blue line shows issues in baseline.

blue line in Figure 1 shows the baseline for nasta'liq and naskh style. The Figure shows that the baseline features are different for nasta'liq and naskh.

We present a novel baseline estimation method for only ghost character for online input. The proposed approach is divided into three phases:

Phase I: Separation of diacritical marks.
Phase II: Primary baseline estimation.
Phase III: Locally baseline estimation.

Phase I: Separation of diacritical marks

As in literatures, the diacritical marks created problem for baseline estimation. Thus in the first step the secondary strokes that is, diacritical marks are segmented based on the size of the strokes shown in Figure 3(c). These ghost characters are used for baseline estimation. The width and height of each stroke is calculated if it is less than threshold value then it is considered as secondary strokes. Although some of the secondary strokes are of greater size, this issue is resolved during the locally baseline estimation. The secondary strokes are discarded and primary strokes are forward for baseline estimation.

Phase II: Primary baseline estimation

For primary baseline estimation we used projection based method. The primary baseline estimation is only to find a rough baseline for locally baseline estimation. The horizontal projection based method counts the number of elements on horizontal line. The maximum number of elements on horizontal line is the baseline. As the secondary strokes are removed during the phase I, thus projection baseline is the estimated baseline on ghost character and gives good result by eliminating the influence of diacritical marks on baseline estimation proposed in literature.

Phase III: Locally baseline estimation

Although projection based baseline is robust and it is very easy to estimate while this method requires a long straight line of text. Whereas in the case of handwritten text especially for online handwritten the length of line/words may be very short or it is very difficult to find a single baseline due to large variation in handwritten text. In the third phase some features are extracted that helps in baseline detection. These features are used to estimate the baseline locally with the help of primary baseline. The features and local baseline estimation is fully dependent on the style of script that is, Nasta'liq has different set of features with different rules as compared to Naskh. Figure 3 describes the proposed baseline extraction method.

Features lying on the baseline are extracted that is, ray, bey etc shown in Figure 4. As in Nasta'liq the last character lies on the baseline, thus for Nastaliq the last shapes of the character is extracted for locally baseline estimation. Whereas for Nask font, local baseline project tion is used with some additional features. The baseline is estimated little above the baseline extracted using features for Naskh.

For primary baseline estimation α is computed based on horizontal projection as shown in Figure 3(c) on ghost character. The primary baseline estimation is used to

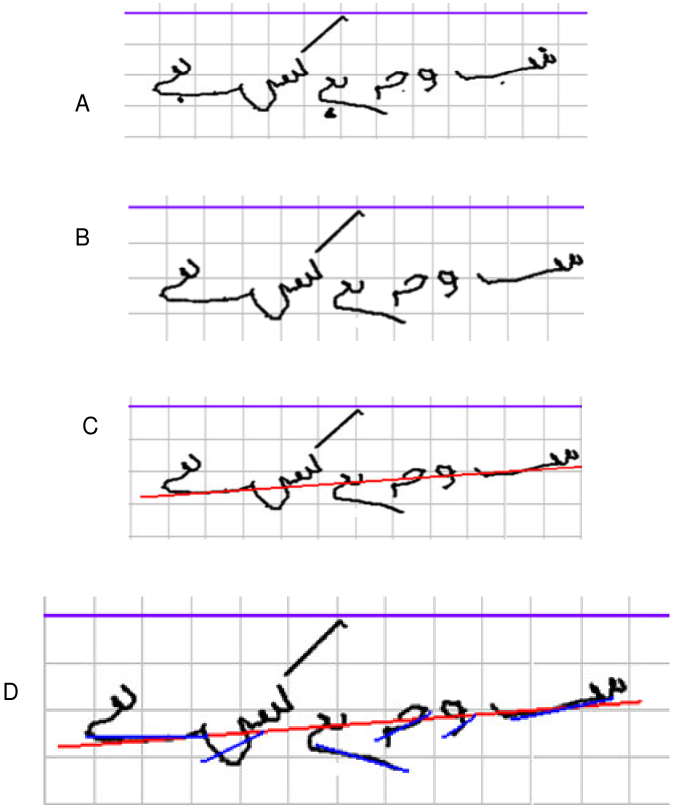

Figure 3. (A) Raw input strokes. (B) Ghost shapes after separation of secondary strokes. (C) Primary baseline estimation based on projection. (D) Locally baseline estimation based on features.

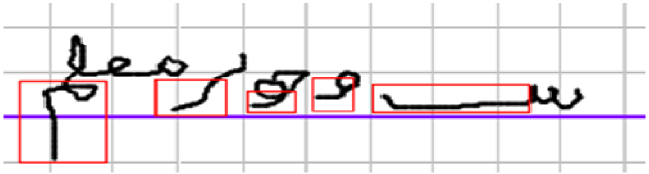

Figure 4. Features for baseline estimation.

reduce the error which occurs by using the feature based approach. Then the second baseline is computed based on the features and pre baseline. The role of primary baseline is to compute the exact angle for local baseline by using the following relation.

For each ligature

If $|\alpha - \beta i| < \theta$ then estimated angle βi else estimated angle α

Where β is locally computed angle of each ligature.

The skewness is performed on both ghost strokes and secondary strokes. The angle of secondary strokes is the same as the angle of associated ghost stroke and it is performed using the following relation.

For ghost strokes

$x' = x \cos \beta i - y \sin \beta i$

$y' = x \sin \beta i + y \cos \beta i$

For secondary strokes

For All secondary strokes associated to ith ghost ligature.

$x' = x \cos \beta i - y \sin \beta i$

$y' = x \sin \beta i + y \cos \beta i$

CONCLUSION

Baseline estimation is one of the most difficult and important phase in character recognition and it has direct influence accuracy. Due to the complex nature of the Urdu scripts based languages, baseline estimation for handwritten text is a very difficult task. We present a novel technique for baseline estimation for cursive handwritten Urdu script written in Nasta'liq and Naskh styles. Firstly the secondary strokes are segmented form the raw input strokes. Then primary baseline is extracted using the horizontal projection on ghost shapes. Finally the local baseline of each ligature is estimated based on features and primary baseline estimation. The presented approach gave good result due to mixture of local baseline estimation over global baseline estimation and reduction of diacritical marks. The proposed method provides accuracy of about 80.3 and 91.7% for Nasta'liq and Naskh font respectively.

REFERENCES

Boubaker H, Kherallah M, Alimi AM (2009). New Algorithm of Straight or Curved Baseline Detection for Short Arabic Handwritten writing, 10th International Conference on Document Analysis and Recognition.

Pechwitz M, Maergner V (2003). "HMM based approach for handwritten Arabic word recognition using the IFN/ENITdatabase," in Proceedings of the International Conference on Document Analysis and Recognition (ICDAR '03), Edinburgh, Scotland, August 2003, pp. 890-894.

Maddouri SS, Abed HE (2008). "Baseline Extraction: Comparison of Six Methods on IFN/ENIT Database " Intentional Conference on Frontiers in Handwritten Recognition.

Faisal S, Ul Hasan A, Dainel K, Thomas M (2004). Breuel, "Layout Analysis of Urdu Document Images" Workshop on Computational Approaches to Arabic Script-based Languages.

Combining Slanted-Frame Classifiers for Improved HMM-Based Arabic Handwriting Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(7).

Jawad H, AlKhateeb, Jinchang Ren, Stan SI, Jianmin J (2008).

Knowledge-based Baseline Detection and Optimal Thresholding for Words Segmentation in Efficient Pre-processing of Handwritten Arabic Text, Fifth International Conference on Information Technology: New Generations, pp. 1158-1159.

Farooq F, Govindaraju V, Perrone M (2005). "Preprocessing methods for handwritten Arabic documents". ICDAR (Proceedings of the Eighth International Conference on Document Analysis and Recognition), pp. 267-271.

Benoureth A, Ennaji A, Sellami M (2008). "Semi-Continuous HMMs with Explicit State Duration Applied to Arabic Handwritten Word Recognition, Pattern Recognition Letters, 29(12): 1742-1752.

Muhammad IR, 1Fareeha A, Husain SA, Abdel B, Muhammad S (2010). HMM and Fuzzy Logic: A Hybrid Approach for Online Urdu Script Based Languages Character Recognition, Int. J. Comp. Mathematics, Submitted. doi:10.1016/j.knosys.2010.06.007.

Ben NA, Bouslama F (2003). “Classification of Arabic Script Using Multiple Sources of Information: State of the Art and Perspective,” Int’l J. Document Anal. Recog., 5: 195-212.