A Framework for Word Spotting In Scanned Urdu Documents by Exploiting the Dot Orientation

Muhammad Shafi[1], Faisal Iqbal[2], Iftikhar Ahmed Khan[1], Muhammad Irfan Khattak[3], Mohammad Saleem[4], Naeem Khan[3]

[1]Department of Computer Software Engineering, University of Engineering and Technology Peshawar, Pakistan
[2]Department of Telecommunication Engineering, University of Engineering and Technology Peshawar, Pakistan
[3]Department Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan
[4]Department of Computer Science, CECOS University, Peshawar, Pakistan
shafi@nwfpuet.edu.pk

Abstract: Urdu is one of the most widely used languages in the world and there is a need of developing character recognition and word-spotting algorithms, so that Urdu literature can be made easily accessible and searchable to the Urdu reading population. Although there has been a sizeable research for character recognition, very few articles have been published for word-spotting in Urdu language. Unlike English language (with only two alphabets with dots), in Urdu language 17 out of 38 alphabets have dots either above or beneath them. This paper presents a data reduction framework, based on exploiting the dot orientation for word spotting in Urdu scanned documents. After applying the proposed scheme, the number of eligible candidates for the target word is greatly reduced. As demonstrated in the Results and Analysis section, the proposed algorithm has shown promising results with an average data reduction rate of 79.8%.

[Muhammad Shafi, Faisal Iqbal, Iftikhar Ahmed Khan, Muhammad Irfan Khattak, Mohammad Saleem, Naeem Khan. **A Framework for Word Spotting In Scanned Urdu Documents by Exploiting the Dot Orientation.** *Life Sci J* 2013; 10(7s): 1163-1171]. (ISSN: 1097-8135). http://www.lifesciencesite.com 185

Keywords: word spotting, tilt removal, horizontal profiles, dot spotting

1. Introduction

The ability of a computer to understand hand-written and scanned documents is important as it can be utilized in efficient searching, data mining, computer intelligence and text pattern analysis etc. Character recognition has been a very active research topic in the computer vision community for a long time. English, being the most widely spoken language of the world, has got much attention of researchers for character recognition. Some work has also been carried out by researchers for character recognition in Urdu language. Text search in a scanned document is a sub topic of character recognition. In case of word spotting, the whole document need not to be recognized by the computer and only the words that have some similarities with the input words in terms of length, shape or some other feature need to be processed. Thus word spotting is less expensive computationally and more feasible as compared to optical character recognition (OCR) (Manmatha et al., 1996a).

Urdu is the national language of Pakistan and is spoken in more than 22 countries, with almost 60 million native speakers (Lewis, 2009). Urdu is also one of the two official languages in Pakistan and one of the 23 official languages in India . Unlike English, Urdu is written from right to left (RTL) with different scripts . The most popular and widely used script of writing Urdu language is Nastaleeq, developed from two different scripts i.e. Naskh and Taleeq. Nastaleeq has a problem of overlapping words (Wali and Rehman, 2007), which makes the word spotting in Urdu language harder as compared to English.

Urdu language has 38 alphabets compared to 26 alphabets of English. Out of the 38 alphabets, 17 alphabets have dots above or below them; known as diacritics, as shown in Figure 1. The position and number of dots are used to differentiate an alphabet from other similar shaped alphabets, as shown in Figure 2. Unlike English (with very few dots), the dots can be utilized in Urdu for word-spotting. This paper aims to develop automatic word spotting utility in scanned Urdu documents, exploiting dot orientation.

Figure 1. Comparison of Urdu and English alphabets

In Urdu language, characters are combined to form compound words and/or partial words as shown in Figure 3. For this purpose, the document is first converted to partial words (a connected region with optional number of dots above or below it), based on the connected components; the spotting algorithm is then applied to these partial words.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work, Section 3 explains the proposed algorithm, Section 4 demonstrates the results and the Section 5 concludes the paper with future research directions.

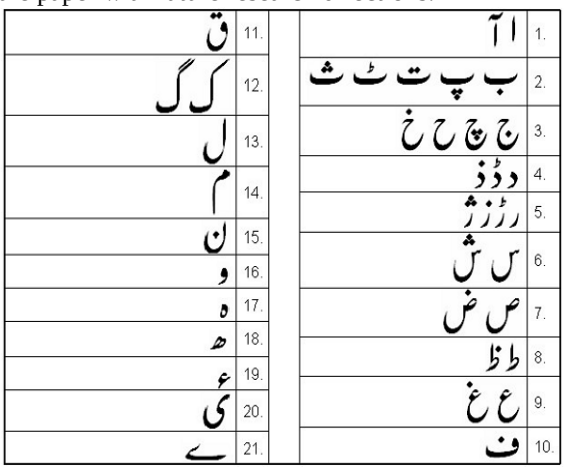

Figure 2. Alphabets with same shape differentiated by dot orientation

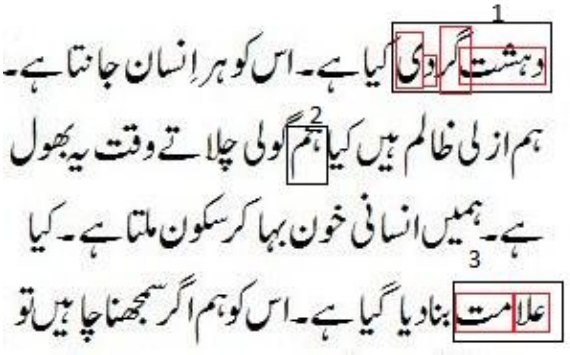

Figure 3. 1 compound word made from five partial words, 2 compound words with no partial words, and 3 compound words with two partial words

2. Related Work

Initial research in word spotting was focused on speech messages (C.S.MYERS et al., 1981, Rose and Paul, 1990, Knill and Young, 1994). It was extended to typed text documents (Kuo and Agazzi, 1994, Chen et al., 1993, Cho and Kim, 2004) and then further to hand-written documents. There are two main types of word-spotting algorithms: query-by-string and query-by-example. In query-by-example systems, character models are trained. At query time these character models are concatenated to form a word model. The next step in the models is the calculation of probability of each word. One of the advantages of these systems is that they allow searching any keyword once the character model is trained. Query-by-string systems are very similar to optical character recognition. Although they are not supposed to obtain the whole transcript, each and every character of the document is processed. Thus these systems have almost all the drawbacks of character recognition including high computational cost and processing time (Rodr ́guez-Serrano and Perronnin, 2009).

In query-by example systems, the input is a small image of a word and the outputs are also images containing words that are most similar in appearance to the query word. So it is quite similar to content-based image retrieval. In these systems, similarity of input word with the target words is calculated by evaluating the distance measure. These systems are specifically good for word-spotting and quite fast because all the characters in the document need not to be processed, and only a limited number of word-images similar in appearance to the query word-image go through all the processing steps. So it is not necessary to decode every single character and word in the document in order to perform word search. Query-by-example systems are further classified into two groups: holistic and local features systems.

In holistic approach, each of the word-images is described with a single feature vector. The distance is calculated between the feature vector of the query image and the target images to do word-spotting. Manmatha et al., (1996b) used pixels of the image as feature vector and applied the Scot and Longuet-Higgens (SLH) distance (Scott and Longuet-Higgins, 1991), which is invariant to affine transformations, to do word spotting. Zhang et.al used Gradient Structural Concavity (GSC) binary features and then used correlations for words matching. In the local-feature approach, the word-image is described as local features. For example, Leydeir et al. (Leydier et al., 2005) used gradient angles as local features and cohesive elastic distance for matching. Rothfeder et al., (2003) used corner detectors as local features and elastic distance between corners as the matching strategy. The state-of-the-art technique for distance measurements between the features is dynamic time warping (DTW) (Berndt and Clifford, 1994). DTW has shown relatively better accuracy as compared to classical feature matching techniques. Some of the local features that have been used in the literature are word-profiles (Kolcz et al., 2000, Rath and Manmatha, 2003b), Eigenslits (Terasawa et al., 2005) and contours (Adamek et al., 2007). Rath and

Manmatha (2003a) have provided a comprehensive comparison of different features: projection profile, partial projection profile, upper/lower word profile, background to Ink transitions, grayscale variance, Gaussian smoothing and Gaussian derivatives, for word spotting in historical documents using DTW matching technique.

Since the research is investigating word-spotting instead of recognition of the whole document, Query-by example systems with customization to Urdu language will be used. To search the documents, dots will be exploited. Only the words that are similar in appearance i.e. that have similar dot orientation to the query word are targeted. Figures 4, 5 and 6 shows that limited numbers of words are targeted in these methods.

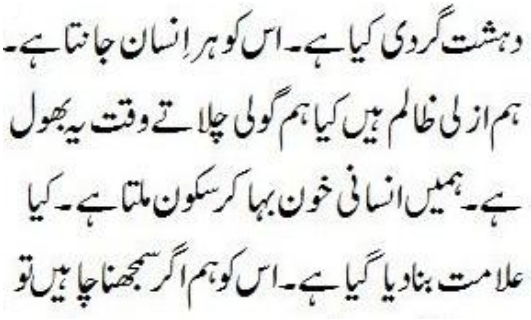
دہشت گردی کیا ہے۔اس کو ہر انسان جانتا ہے۔
ہم ازلی ظالم ہیں کیا ہم گولی چلاتے وقت یہ بھول
ہے۔ہمیں انسانی خون بہا کر سکون ملتا ہے۔کیا
علامت بنادیا گیا ہے۔اس کو ہم اگر سمجھنا چاہیں تو

Figure 4. Target Image

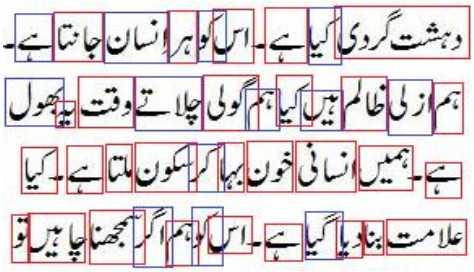

Figure 5. Individual Word-Images

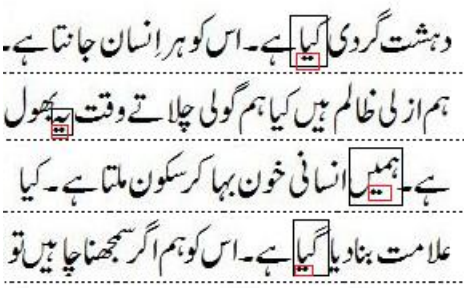

Figure 6. Possible target words (based on dots) for "ye" as query word

Several Research papers have been published for designing OCR for Urdu Documents (Fareen et al., 2012) but very few articles have been published for word spotting in printed Urdu scanned documents. The following paragraph summarizes the closely related work to Urdu word-spotting from the literature.

Abidi et al. (2011) presented a word-spotting mechanism for scanned Urdu documents. The mechanism is based on partial word features. In their proposed method, the image is binarized and partial words are extracted first, followed by a retrieval step in which words in the query are compared (based on features) with the indexed partial words. For feature extraction, two scalar approaches: aspect ratio, convex area and four vector approaches i.e. upper profile, lower profile, ink to non-ink trace and vertical projection were used. The retrieval is then based on smart sorting, DTW and measuring Relative distance and combining. Sagheer et al. (2010a) and Sagheer et al. (2010b) presented a word-spotting algorithm based on connected components integrated with diacritics. A sliding window of four was used for generating a candidate word. Gradient and features of candidate words were compared for recognition. For verification and rejection, sum based classifier was used which compared features like dots and aspect ratio number of black pixels. Pal and Sarkar (2003) used a combination of topological, contour and water reservoir concept based features for word-spotting. Dots, punctuation marks, small modifiers etc. are filtered out first. Hough transform is used for finding skew angle while horizontal profiles are used for line separation. Wshah et al. (2012) used Markov models for word spotting in handwritten Arabic documents. The work done by Wshah et al. (2010) is closely related to the proposed mechanism that also uses the dots orientation for word spotting for Arabic languages. The structural differences of Urdu and Arabic languages and the difference in the processing steps differentiate our proposed method from their method.

3. Proposed Algorithm

The flowing flow chart shown in figure 7 demonstrates the proposed algorithm. The algorithm composed of five main components which are discussed below in details:

3.1 Document Tilt Removal

There is always a chance of tilted at some angle document being scanned. Most of the word-spotting algorithms in literature (and also the proposed method) need the tilt removed from the document first and hence the document to be perfectly aligned with the axis as a pre-requisite.

As discussed in detail in (Hull, 1998), the document skew angle removal methods could be broadly classified into four categories: horizontal projection profile, Hough transform, Feature location

distribution and directionally sensitive masks response distribution.

The horizontal projection profile method has been incorporated in this research for skew angle removal due its simplicity and ease of calculations. Horizontal projection profile of a document is an array of size equal to the number of pixel rows in a document. Each bin of this array is populated by adding all the intensity levels of the particular row in the document. The document is rotated at various angles (with small variation in the angle size) and the horizontal projection profile is calculated for each rotated angle. The angle that gives the most regular-shape and peaks in the horizontal projection profile is considered as the skew angle.

Figure 8 shows a document rotated at various angles and corresponding horizontal projection profiles. As shown in the figure, the profile is much regular shaped with highest frequency and amplitude for the skew angle of 0 degrees, suggesting that there was no tilt in this document.

3.2 Text Size Variations

Since the dots spotting section is based on the size of the dots in the document, it will not be applicable to a document containing text in different font sizes. The parts of the documents with different font sizes should be separated first. The horizontal profiles of the document obtained in section 3.1 could be easily utilized for this purpose as shown in Figure 9. Here it is assumed that a single line of text doesn't contain characters of different sizes. After getting the horizontal profiles, the height of each black strip is used to predict the size of characters in that particular line of text. All the lines of text with same size are kept in one document and hence several separate documents are then created based on the size of text. Each of these documents is then separately fed to the dots spotting algorithm of the next section

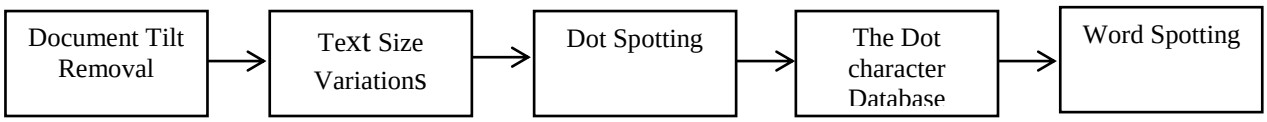

Figure 7. Flow chart for the proposed algorithm

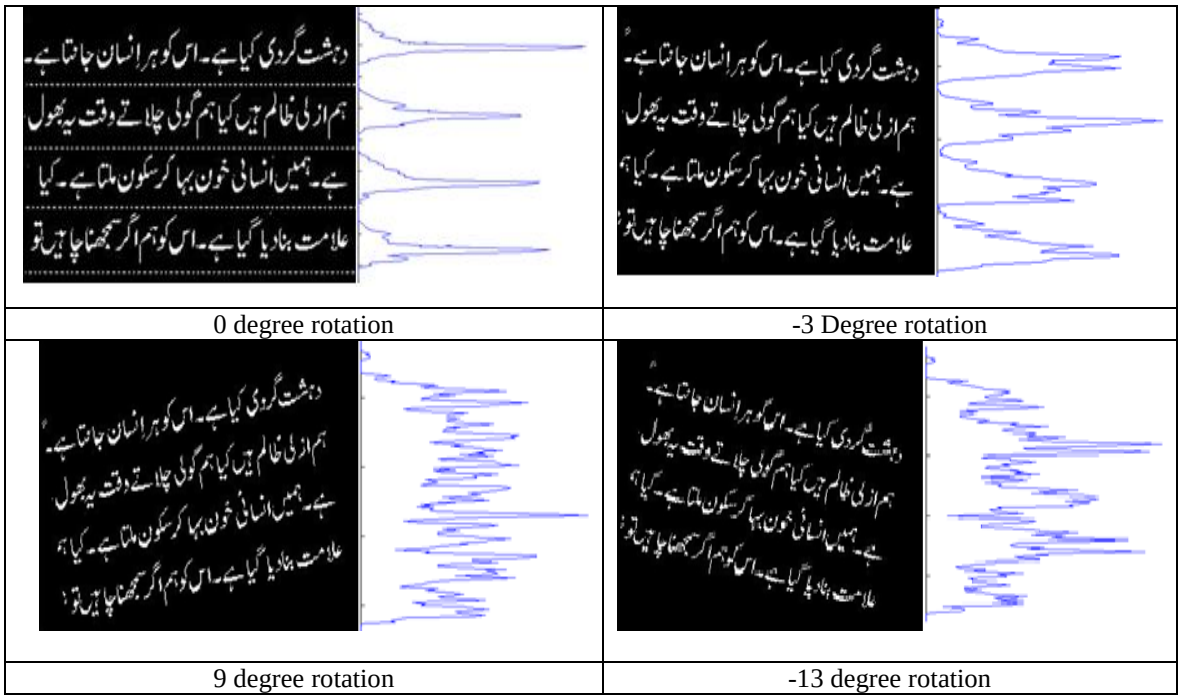

Figure 8. Image rotated at different angles with horizontal projection profiles

3.3 Dots Spotting

It has been observed that in Urdu text, single dots are the only connected regions that have almost circular shapes. The circularity of all other connected regions is less than a single dot. This maximum circularity property of a dot is used in the proposed scheme for dots extraction in a scanned Urdu document. Some false positives are further removed by using the presumption that all dots in a text document (that contains text of the same size) have

almost the same area of the dots as shown in Figure 10. So any circular connected region with a size much larger or smaller than the average dot-candidates are excluded from the dots-candidates. Double dots are then detected using the following heuristics.

- Size of the connected region is almost double to that of a single dot.
- The orientation is horizontal.

Figure 9. Horizontal profiles with different font size

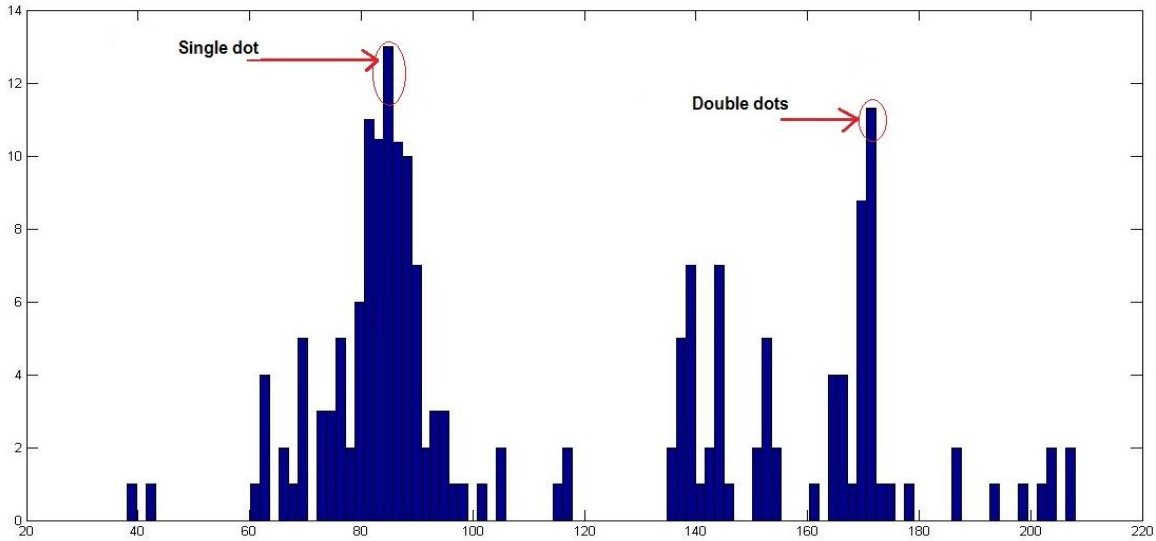

Figure 10. Histogram showing the size of dots on x-axis and number of dots in a document on y-axis

3.4 The Character-Dot Orientation Database

The horizontal projection profiles, calculated in section 3.2 are used to separate each line of text. The upper left and lower-right are located for each connected region in the document, where the single and double dots have already been removed. Each connected regions are then ornamented with these attributes: upper-left corner, lower-right corner, number of dots below the region and number of dots above the region. The dots below and above are calculated based on the connected region bounding-box, dots positions and individual text lines separation. The sentence-ending dots and other punctuations are removed based on the horizontal projection profiles and sizes respectively.

A simple database with the table structure shown in Table 1 is then created to store all these connected region annotations, and make the word-searching easy and fast (Further details can be found in section 3.5).

Related information stored Table 1 below contains ID, Upper-left corner, Lower-right corner, dots above, and dots below.

Table 1. Database with connected annotations

ID	Upper-Left Corner	Lower-right Corner	Dots above	Dots below
15	12,30	17,34	2	0
16	20,31	17, 35	0	2
17	67,23	64,26	0	3
18	73,78	74,85	0	0

3.5 Word Spotting

Instead of searching the whole word, partial words are searched in the connected regions database created in the previous section. It has been observed

that usually there are fewer partial words with more dots, than those with few or no dots. So, if the word (that is being searched) comprises multiple partial words, the partial word with most dots is searched first, followed by others in order of number of dots. For example, the word Pakistan will be searched as follows:

Step 1

Divide the word in partial words i.e. Pa, kista and Noon. This is shown in Figure 11.

Step 2

Search for the partial word "Pa" (having the highest number of dots) first in the database shown in figure 12; for which there are only five eligible candidates.

Step 3

The next partial word to be searched is "Kista". So all the characters returned in step 2 are searched for a partial word to their left with two dots above. All other candidates are removed shown in figure 13.

Step 4

The final partial word to be searched is "noon". All the candidates returned in step 3 are further searched for a partial word towards left with a dot above. All other partial words are removed.

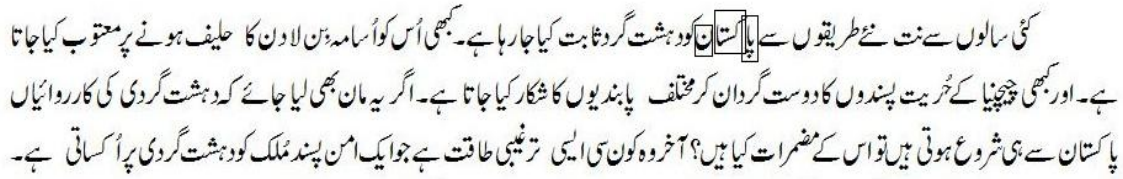

Figure 11. Indexed word having three partial words

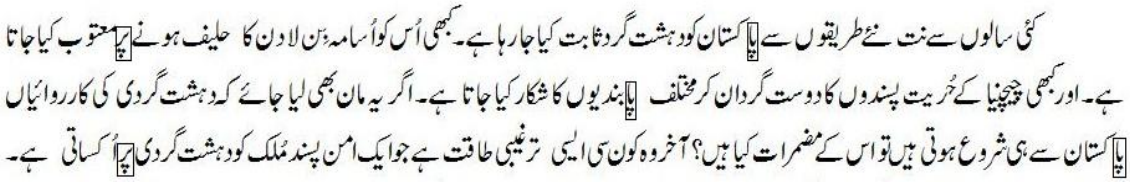

Figure 12. Eligible candidates for the Pa i.e. having 3 dots below

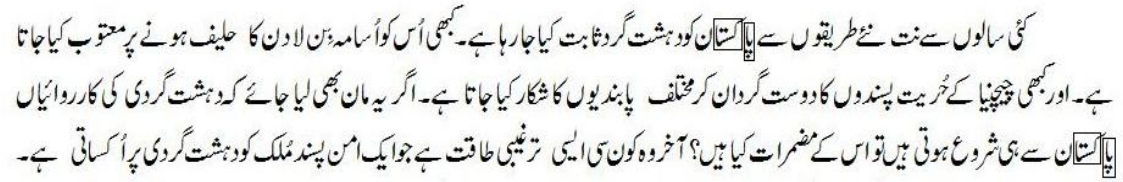

Figure 13. Search result for partial word at the left side of "pa" having two dots above

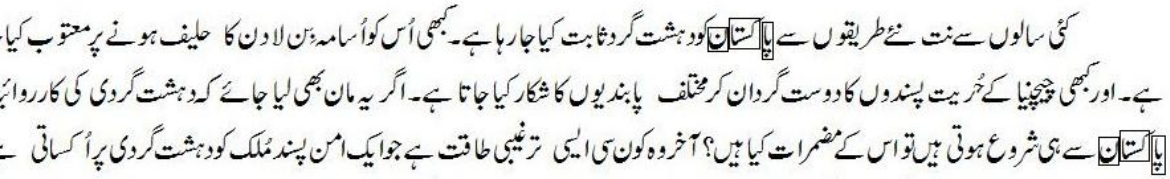

Figure 14. Search for partial word at the left of "pakista" having single dot above

4. Result and Analysis

The proposed algorithm was applied to different documents and results were generated. Following is statistical analysis for a sample document, after applying the proposed scheme.

The document has 126 words and 215 partial words. The partial words constitute 95 unique words.

The average eligible candidates for each partial word are 29.4. i.e.13.4 %; it means that 86.4% of partial words were removed.

Table 2 Statistical analysis for selected document

Word	Eligible candidates	Percentage
Jab	17	13.49206
Sirf	8	6.349206
Khwab	6	4.761905
Hi	110	87.30159
Ajenda	2	1.587302
Hoga	30	23.80952
To	14	11.11111
Phir	1	0.793651
Apni	2	1.587302
Dewarain	1	0.793651

Bhi	19	15.07937
Aisa	9	7.142857
Salook	46	36.50794
Karti	8	6.349206
Hainn	17	13.49206
Yaad	9	7.142857
Aya	7	5.555556
Keh	110	87.30159
Kashmiri	1	0.793651
Rehnuma	8	6.349206
Janab	1	0.793651
yassenmalik	1	0.793651
Ney	19	15.07937
Jo	19	15.07937
Imran	4	3.174603
Khan	6	4.761905
Sahib	14	11.11111
Ki	110	87.30159
Main	17	13.49206
Aine	8	6.349206
Ke	110	87.30159
kaha	110	87.30159
lekin	3	2.380952
Aik	9	7.142857
hum	110	87.30159
hazar	4	3.174603
saal	30	23.80952
tak	14	11.11111
jhung	2	1.587302
larain	9	7.142857
Ge	110	87.30159
Us	30	23.80952
waqt	1	0.793651
zamini	1	0.793651
haqaeeq	1	0.793651
aaj	8	6.349206
Se	110	87.30159
bilkul	12	9.52381
mukhtalif	1	0.793651
the	14	11.11111
bharat	1	0.793651
Ka	110	87.30159
ghumand	19	15.07937
torney	1	0.793651
Ke	110	87.30159
liye	110	87.30159
awam	21	16.66667
Ka	110	87.30159
moral	14	11.11111
bharhar	3	2.380952
tha	14	11.11111
baad	19	15.07937
kaha	110	87.30159
tha	14	11.11111
nehru	12	9.52381
cricket	30	23.80952
ke	110	87.30159
madah	21	16.66667
rahe	30	23.80952
siyasi	9	7.142857
maidan	2	1.587302
barisagheer	2	1.587302
nichely	4	3.174603
tabqat	1	0.793651
siyasat	1	0.793651
ghandi	5	3.968254
aur	21	16.66667
se	110	87.30159
ziyada	1	0.793651
bhutto	19	15.07937
asarat	1	0.793651
martab	1	0.793651
huwe	30	23.80952
unhon	5	3.968254
darasal	14	11.11111
apney	2	1.587302
pasandeeda	1	0.793651
khiladi	21	16.66667
ko	110	87.30159
rasta	4	3.174603
chunney	1	0.793651
ko	110	87.30159

Apart from the selected sample document, a total of 8714 partial words and 3615 unique words were randomly picked from different documents and analyzed. The following table contains the statistics.

Table 3. Result and analysis

Number of words	Total partial words	%age of Average Eligible candidate for each partial word	Average Percentage reduction
3615	8714	20.2 %	79.8%

5. Conclusions and Future Work

This paper presented a word-spotting algorithm for scanned Urdu documents, exploiting the dots orientations. Due to the higher number of dots in Urdu alphabets (as compared to English) and the ease of calculation, the dots orientation was proved to be a good choice for word-spotting, which was demonstrated in the results and analysis section. The dots orientation combined with other techniques like contour matching, water reservoir method and Markov models could potentially demonstrate more accurate word-spotting mechanisms. Neural networks can also be trained for word spotting. All of these techniques will be further investigated in future.

Corresponding Author:
Dr. Muhammad Shafi,
Department of Computer Software Engineering,

University of Engineering and Technology, Peshawar, Pakistan.
E-mail: shafi@nwfpuet.edu.pk

References

[1] Urdu [Online]. Available: http://www.ethnologue.com/language/URD.
[2] Urdu (اردو) [Online]. Available: http://www.omniglot.com/writing/urdu.htm.
[3] Urdu Computing Information (Penn State) [Online]. Available: http://tlt.its.psu.edu/suggestions/international/bylanguage/urdu.html#script.
[4] Abidi, A., Siddiqi, I. & Khurshid, K. Towards Searchable Digital Urdu Libraries - A Word Spotting Based Retrieval Approach. 2011 International Conference on Document Analysis and Recognition (ICDAR), 2011. 1344-1348.
[5] Adamek, T., Connor, N. E. & Smeaton, A. F. 2007. Word matching using single closed contours for indexing handwritten historical documents. International Journal on Document. Analysis and Recognition 9, 153–165.
[6] Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. KDD workshop, 1994. Seattle, WA, 359-370.
[7] C.S.Myers, Rabiner, L. R. & A.E.Rosenberg 1981. On the use of dynamic time warping for word spotting and connected word recognition Bell System Tech. J. 60
[8] Chen, F. R., Wilcox, L. D. & Bloomberg, D. S. 1993. Word spotting in scanned images using hidden Markov models. IEEE Conference on Audio, Speech and Signal Processing.
[9] Cho, B. J. & Kim, J. H. 2004. Print keyword spotting with dynamically synthesized pseudo 2d HMMs. Pattern Recognition 25 999-1011.
[10] Fareen, N., Khan, M. A. & Durrani, A. Survey of Urdu OCR: An Offline Approach. Conference on Language and Technology 2012 (CLT12), 2012 Lahore,Pakistan. 67-72.
[11] Hull, J. J. 1998. Document image skew detection: Survey and annotated bibliography. Series In Machine Perception And Artificial Intelligence, 29, 40-66.
[12] Knill, K. M. & Young, S. J. 1994. Speaker dependent keyword spotting for accessing stored speech. Technical Report CUED/F-INFENG/TR 193. Cambridge University Engineering Department.
[13] Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R. & Popescu, G. V. 2000. A line-oriented approach to word spotting in handwritten documents. Pattern Anal. Appl, 3, 153–168.
[14] Kuo, S. S. & Agazzi, O. E. 1994. Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models. IEEE Trans. Pattern Anal. Mach. Intell., 16, 842–848.
[15] Lewis, M. P. 2009. Ethnologue: Languages of the World. 16 ed.: SIL International.
[16] Leydier, Y., Bourgeois, F. L. & Emptoz, H. Omnilingual segmentation-free word spotting for ancient manuscripts indexation. International Conference on Document Analysis and Recognition, 2005. 533-537.
[17] Manmatha, R., Han, C., Riseman, E. & Croft, W. Indexing handwriting using word matching. Proceedings of the first ACM international conference on Digital libraries, 1996a. ACM, 151-159.
[18] Manmatha, R., Han, C. & Riseman, E. M. Word spotting: A new approach to indexing handwriting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1996b. 631.
[19] Pal, U. & Sarkar, A. Recognition of printed Urdu script. Proceedings. Seventh International Conference on Document Analysis and Recognition, 2003., 2003. 1183-1187.
[20] Rath, T. M. & Manmatha, R. Features for word spotting in historical manuscript. 7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2003a.
[21] RATH, T. M. & MANMATHA, R. Word image matching using dynamic time warping. 2003 IEEE Conference on Computer Vision and Pattern Recognition, 2003b. 521–527.
[22] Rodr ́guez-Serrano, J. A. & Perronnin, F. 2009. Handwritten word-spotting using hidden Markov models and universal vocabularies. Pattern Recognition, 42, 2106-2116.
[23] Rose, R. C. & Paul, D. B. 1990. A hidden Markov model based keyword recognition system International Conference on Acoustics, Speech, and Signal Processing.
[24] Rothfeder, J., Feng, S. & Rath, T. 2003. Using corner feature correspondences to rank word images by similarity. Workshop on Document Image Analysis and Retrieval.
[25] Sagheer, M. W., Chun Lei, H., Nobile, N. & Suen, C. Y. Holistic Urdu Handwritten Word Recognition Using Support Vector Machine. Pattern Recognition (ICPR), 20th International Conference on, 23-26 Aug. 2010 2010a. 1900-1903.
[26] Sagheer, M. W., Nobile, N., Chun Lei, H. & Suen, C. Y. A Novel Handwritten Urdu Word Spotting Based on Connected Components Analysis. Pattern Recognition (ICPR), 20th International Conference on, 23-26 Aug. 2010 2010b. 2013-2016.
[27] Scott, G. L. & Longuet-Higgins, H. C. 1991. An

algorithm for associating the features of 2 images. Proceedings of the Royal Society of London Series 244, 21-16.
[28] Terasawa, K., Nagasaki, T. & Kawashima, T. Eigenspace method for text retrieval in historical document images. 8th International Conference on Document Analysis and Recognition, 2005. 436–441.
[29] Wali, A. & Rehman, S. 2007. Implementation of Reverse Chaining Mechanism in Pango for Rendering Nastaliq Script. Second Workshop of Computational Approaches to Arabic Script-based Languages. USA: Stanford University.
[30] Wshah, S., Govindaraju, V., Yanfen, C. & Huiping, L. A Novel Lexicon Reduction Method for Arabic Handwriting Recognition. 20th International Conference on Pattern Recognition (ICPR), 2010. 2865-2868.
[31] Wshah, S., Kumar, G. & Govindaraju, V. Script Independent Word Spotting in Offline Handwritten Documents Based on Hidden Markov Models. 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012 2012. 14-19.

7/4/2013