Towards Searchable Digital Urdu Libraries – A Word Spotting Based Retrieval Approach

Ali Abidi
National University of Sciences & Technology (MCS-NUST)
Islamabad, Pakistan
abidi@mcs.edu.pk

Imran Siddiqi
National University of Sciences & Technology (MCS-NUST)
Islamabad, Pakistan
imran.siddiqi@mcs.edu.pk

Khurram Khurshid
Institute of Space Technology
Islamabad, Pakistan
khurram.khurshid@ist.edu.pk

***Abstract*—Libraries in South Asia hold huge collections of valuable printed documents in Urdu and it is of interest to digitize these collections to make them more accessible. The unavailability of an OCR for Urdu however limits the concept of a digital Urdu library to scanning of documents only, offering very limited search facility based on manually assigned tags. We address this issue by proposing a word spotting based keyword search method for information retrieval in digitized collections of printed Urdu documents. The proposed method is based on segmentation of Urdu text in to partial words and representing each partial word by a set of features. To search a specific word (or phrase), the user provides a query in the form of an image. Comparing the features of the partial words in the query image with the ones already indexed, the user is provided with a list of documents containing occurrences of the queried word. The system evaluated on 50 Urdu documents exhibited a recall of 95.17% and a precision of 94.3%.**

Keywords – Urdu digital libraries, Word Spotting, Dynamic Time Warping

I. INTRODUCTION

Libraries across the world contain valuable information in various forms, the printed form being the most common. Searching these printed items for some specific information could be a very time consuming job. The advent of digital libraries has been a revolutionary milestone in information retrieval [1]. Scanned images of printed documents can be consulted more efficiently and in more convenient manner. Unfortunately, the job is only half done if the huge volume of scanned library content is not searchable. Availability of high performance Optical Character Recognition (OCR) systems has addressed this issue to a significant extent making it possible to retrieve the required information in a span of few seconds.

Despite its revolutionary contributions to indexing and retrieval of digitized documents, OCR does not present a complete solution to the problem with limitations to cope with handwritings and degraded ancient collections. In addition, there are a number of scripts for which the OCR technology is either non-existent or is still in its infancy. For all these reasons, word spotting presents an attractive alternative to traditional OCR for indexing and retrieval of digitized document collections. Our work in this domain focuses on printed Urdu documents and we present a system for information search by spotting the instances of a given query word in collections of Urdu documents.

The proposed system allows searching a query word image in the collection of indexed documents. A set of features is extracted from each partial word of a given word and the documents are indexed. In the retrieval phase, a three-stage feature matching approach is used to spot the instances similar to the query word.

This paper is outlined as follows. We first give a brief account of some of the well-known word spotting methods followed by some challenges associated with the Urdu language. We next describe the proposed word spotting methodology and the experimental results and finally we give a conclusion and some end remarks

II. RELATED WORK

Information retrieval has been one of the most addressed research areas during the last decade. The inception of digital libraries has driven the development of efficient indexing and retrieval systems providing access to the requested information at a click. One of the revolutionary developments in this area was the availability of OCR systems. However, due to certain limitations (as already discussed), word spotting emerged as an appealing substitute to OCR. While OCR converts the text into machine readable format, word spotting relies on matching the shapes of words without recognizing the words to be matched.

The document recognition community has proposed a large number of word spotting techniques over the years. These are generally categorized into two main classes; image based and feature based matching techniques [2]. Another classification is to divide these methods into segmentation based or segmentation free techniques as described by [3-6]. The two categories are popularly known as analytical and holistic techniques respectively.

Holistic techniques [6-10] view the text word image as a unit that cannot be further segmented. These techniques are based on the principle that characters are easily recognized if they appear within a word rather than in isolation [11]. Typically words are represented by profile based feature sequences and matched with the query word.

Analytical techniques [12-16] segment a text image into smallest possible independently recognizable units. Typically

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.270

segmentation is carried out at character [12] or connected component [13] level. Features based on sliding windows have also been effectively used in [14, 15].

The only work that we could find on Urdu text is proposed in [17]. The method however is not word spotting in its true sense and is based on a word recognition system using compound features and SVM classification. A sliding window of four connected components (CCs) is used to generate candidate words with the assumption that a word comprises a maximum of four components. The system reports 70% recall and 51% precision on the CENPARMI Urdu database.

III. Challenges – Urdu Language

Urdu is the national language of Pakistan and a major language of India with speakers all over the world. Urdu originated from various languages with most pronounced effects of Arabic and Persian. Like both of these languages, Urdu is also written from right to left and word formation is done by combining various Partial Words (PWs) where a PW is made up of various combinations of basic characters. Due to the inherent cursive nature of its script and difficult word formation approach, Urdu becomes a very challenging language to deal with as discussed in the following.

A. *Appearance of Characters Within Words*

Unlike the languages based on Latin alphabet, each of the 39 basic characters in Urdu language has an appearance that depends upon its position within the word. Appearance of same character varies depending upon whether it appears in isolation or as a part of a PW as well as its position (start, middle or end) within the PW, making its segmentation/recognition difficult. A subset of Urdu alphabets and their different appearances are illustrated in Figure 1.

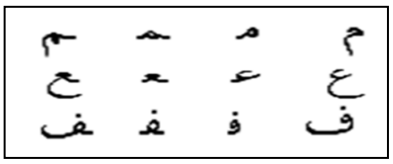

Figure 1. Examples of few Urdu alpahabets having different appearances: isolation, start, middle and end – (Right-to-Left)

B. *Word Segmentation*

The distribution of intra and inter word distances in Urdu is highly non-uniform (Figure 2). Intra word distances often exceed the inter word distances making it practically impossible for techniques like Run Length Smooth Algorithm (RLSA) to be used for word segmentation.

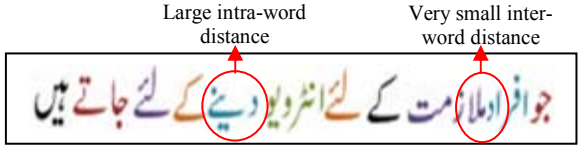

Figure 2. Finding inter word distances is difficult in Urdu script

C. *Character Segmentation*

The highly cursive nature of Urdu script makes the segmentation of PWs into characters very difficult. The contour based segmentation schemes proposed for the Latin alphabet fail once applied to Urdu text. Figure 3 shows an ideal segmentation of Urdu text into basic characters which cannot be achieved with traditional segmentation methods employed in segmentation of Latin text.

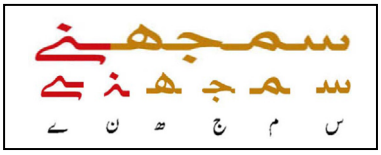

Figure 3. An ideal segmentation of Urdu words into characters

D. *Overlapping of PWs*

Urdu text contains excessive overlapping of adjacent PWs both within a word and between different words. This overlapping adds to the difficulties in the correct segmentation of Urdu text as indicated in Figure 4.

Figure 4. Pronounced overlapping effects in Urdu text

E. *Dots and Diacritic Marks*

Urdu script has the most number of dots and diacritic marks as compared to any other known script. There are a total of 23 characters having these marks (Figure 5). The excessive quantity of these marks makes Urdu language more complicated. Various combinations of dots varying from one to three are used to represent a character and these dots may appear over and underneath the parent character.

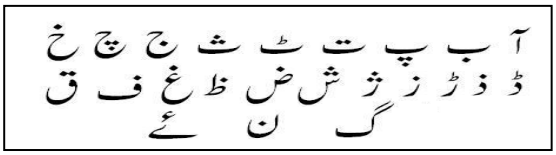

Figure 5. Urdu alphabets having dots and diacritic marks

After having discussed the challenges associated with Urdu text, we present the proposed methodology in the next section.

IV. Proposed Methodology

The proposed word spotting system for Urdu documents is mainly divided into two main parts; indexing and retrieval. During the indexing phase, the document image is binarized and the PWs in the image are extracted. Each PW is represented by a set of (scalar and vector) features. In the retrieval phase, the PWs in the query word image are compared with the PWs in the indexed documents using a three stage matching. These steps are discussed in detail in the following.

A. Text Document Indexing

We first carry out a binarization of the document to separate text from background. Since our research focuses on scanned images of contemporary Urdu books, the images are not very noisy making it possible to use a global thresholding. In our system, we have employed the well-known Otsu's algorithm to extract text in a document. Once the text is binarized, we proceed to segmentation and feature extraction.

1) Text Segmentation: Text segmentation is the most critical step in any retrieval system either using OCR or word spotting. A number of methods for segmentation at line, word and character level have been proposed in the literature. These, however, cannot be directly applied to Urdu text, especially for word and character segmentation due to issues already discussed.

Taking into account these issues, we have chosen not to go for word or character segmentation. Instead we work on partial words (PWs) where each partial word comprises one or more basic Urdu characters combined together to form a part of a word. From the perspective of implementation, we extract the connected components in the binarized text image which, with a few exceptions, correspond to the PWs. Figure 6 shows some PWs extracted from a line of text.

Figure 6. Some partial words (PWs) in a line of text

Although PWs can be extracted directly from the image, we also segment the text into lines (using traditional horizontal projections) which is helpful in merging PWs into words in the retrieval phase.

2) Feature Extraction: After having extracted independent PWs from the text, we proceed to feature extraction. We have defined two scalar and four vector feature values that are extracted from each PW. The scalar features chosen are (i) aspect ratio of each PW and (ii) its convex area which is defined as the total area of polygon formed around a PW normalized by the total area of its bounding box. These features are used in an initial smart sorting as discussed shortly.

In order to capture the shape of the PW more precisely, vector feature sequences are also extracted from each PW. These features have proven to be effective on word [10] as well as character [16] levels on Latin alphabets. The four vector features investigated in our study include:

Upper profile: the distance of first text pixel from the top of the PW's bounding box in each column normalized by the height of the respective bounding box.

Lower profile: the distance of last text pixel from the top of the bounding box and normalized in a similar fashion.

Ink to non-ink transition: the number of transitions from background to the text in each column normalized by the maximum possible transitions in Urdu script (found to be 4).

Vertical projection: the normalized sum of pixel values in each column.

It should be noted that first three vector profiles are obtained using binary image and the last profile is obtained from a gray scale image. Figure 7 illustrates the four vector feature profiles obtained for two different instances of the same PW.

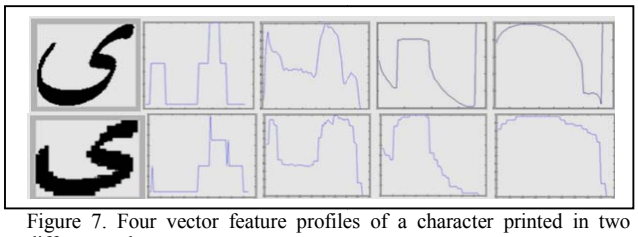

Figure 7. Four vector feature profiles of a character printed in two different styles

Once these features are extracted, an index file is generated for each document in the reference base. For each PW, we keep its position within the image, the two scalar and four vector feature sequences.

B. Retrieval using Word Spotting

During the retrieval phase, the user provides a query word image and expects in return a list of documents containing instances of the queried word. The PWs in the query word are segmented and the same set of features is extracted for each of the PWs. A three-stage matching technique is then employed to spot the instances of query word in the indexed collection of documents. The first two of these serve to match the PWs in the query image with those in the reference base while the last step merges the spotted instances of PWs into words. Each of these steps is discussed in the following.

1) Smart Sorting: Smart sorting is carried out on the basis of scalar features (aspect ratio and convex area) with the objective of filtering out all non-relevant PWs in the reference base. This not only serves to reduce the number of candidate PWs to be compared in the next step but also improves the performance of the system. Only the PWs satisfying the following crieria are passed on to the next matching stage:

$$\frac{R_{query}}{R_{candidate}} \leq T1, \quad 0.9 < T1 < 1.1 \qquad (1)$$

$$\frac{C_{query}}{C_{candidate}} \leq T2, \quad 0.95 < T2 < 1.05 \qquad (2)$$

Where R_{query} and $R_{candidate}$ represent the aspect ratios while C_{query} and $C_{candidate}$ the convex areas of the query and

candidate PWs respectively. The thresholds are chosen so as to allow false positive at this stage but not to miss valid instances of the searched PW. In our evaluations, using smart sorting alone, we filter out about 71% irrelevant PWs leaving only 29% to be matched in the next step.

2) Dynamic Time Warping(DTW): Once the irrelevant PWs are filtered out, the filtered PWs are matched with the query PW using DTW on the four vector features. Among the various matching techniques investigated at word as well as character levels, DTW has been most effective thanks to its ability to account for the nonlinear stretch, style and size of the two profiles to be matched. In our system, we employ DTW matching at PW level comparing the four profiles of the PWs to be matched. By comparing these features using DTW, we are able to cater for changes in style and size.

Once the PWs are spotted, we need to keep only those which are parts of the same word as the query. For this purpose we employ relative distance matching and combining as presented in the following.

3) Relative Distance Matching and Combinining (RDMC): The main objective of this final step is to merge the spotted PWs into complete words and eliminat the irrelevant instances which are spotted correctly but are parts of words other than the query. PWs spotted in a given line of text can only be merged with the PWs in the same line. The simplest merging technique could be to merge all retrieved PWs in a line of text if they appear in the same order as in the query word. This apporach however is highly sensitive to the order of PWs and a sligth change in the order of PWs (for example due to different positioning of dots or presence of noisy components) in query word and retrieved document can result in discarding a correctly retrieved combination of PWs.

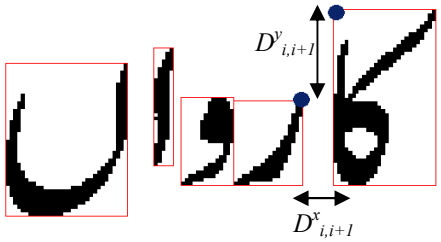

Figure 8. Relative distance calculation in RDMC

For a given query word, RDMC computes the relative distance between all adjacent PWs of the query word along the horizontal and vertical directions. Distances are calculated for adjacent PWs from left to right as:

$$D^x_{i,i+1} = (x^{left}_{i+1} - x^{right}_i)/w_{i,i+1} \quad (3)$$

$$D^y_{i,i+1} = (y^{left}_{i+1} - y^{right}_i)/h_{i,i+1} \quad (4)$$

The numerator in each expression computes the distance between adjacent PWs (*i* and *i+1*) while the denominator represents the normalization by width (height) of the bounding box after merging PWs *i* and *i+1*. The superscripts '*left*' and '*right*' refer to the *x(y)* coordinates of the upper left and right corners of the bounding boxes respectively. The computation of these distances is illustrated in Figure 8.

These relative distances between adjacent pairs of PWs in the query image serve as a reference for combining the retrieved PWs into words. Within each text line of retrieved documents, if the PWs *i* and *i+1* of query word are spotted, they are merged together if the relative (horizontal and vertical) distance between the PWs *i* and *i+1* is comparable to the same distance in the query word:

$$|D^{Q,x}_{i,i+1} - D^{R,x}_{i,i+1}| < T3 \; and \; |D^{Q,y}_{i,i+1} - D^{R,y}_{i,i+1}| < T4 \quad (5)$$

Where $D^Q_{i,i+1}$ and $D^R_{i,i+1}$ represent the relative distance between PWs *i* and *i+1* in query word and retrieved document respectively while *T3* and *T4* are empirically determined constants.

Inspecting each PW spotted in a given line (from left to right) we keep merging the PWs that satisfy the distance constraints defined in (5). As a result of this process we get the words that are similar to the query word as illustrated in Figure 9. Figure 9a shows the PWs retrieved in response to the query word in figure 7. Applying RDMC only the PWs which are part of an occurrence of the query word are retained while the rest are ignored (Figure 9b).

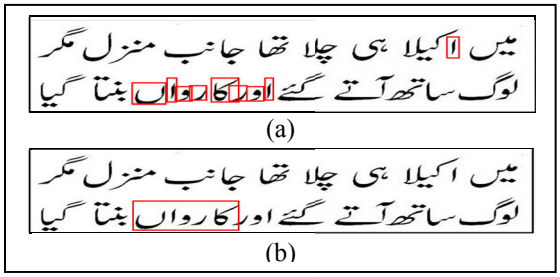

Figure 9. (a) Input to RDMC (b) Formation of relevant candidate word

V. Experimental Evaluation

For the experimental evaluation of our system, we have employed a data set comprising 50 different images taken from 5 different Urdu books. For testing, 65 query words having 435 instances in total were selected. These 65 query words are further divided into 194 PWs having 1654 instances in total. It should be noted that the results of Smart Sorting and DTW are based on PWs only while those of RDMC are based on complete words. The results are summarized in Table 1.

The first step of smart sorting is based on very simple scalar features and naturally the number of false positives is high. The thresholds in smart sorting are also chosen so as not to miss any PWs at this stage. Overall we achieve a reduction

of 71% irrelevant PWs leaving the remaining 29% to be compared in the next step.

The performance of four vector features using DTW is evaluated with and without smart sorting. Without smart sorting, we achieve a precision of 71.52% and recall of 79.87%. Using smart sorting prior to DTW matching not only reduces the number of comparisons and hence the computation time but also improves the precision to 81.23% and recall to 90.51%. Combining the retrieved PWs into words using RDMC, the system reports a recall of 95.17% and precision of 94.3%.

TABLE I. Summary of Results

	Matching Method			
	Smart Sorting	**DTW (With S. Sorting)**	**DTW (W/O S. Sorting)**	**RDMC**
Matching Level	**Partial Words**			**Words**
Query items	1654	1654	1654	435
True Positives	1654	1497	1321	414
False Negatives	0	157	333	21
False Positives	1238	346	526	25
Precision	57.21%	81.23%	71.52%	94.3%
Recall	100%	90.51%	79.87%	95.17%
F-Measure	72.77	85.62	75.46	94.73

VI. CONCLUSION AND PERSPECTIVES

We have presented an effective method for retrieval of printed Urdu documents using word spotting. The proposed methodology relies on segmenting the text into PWs and representing each PW by a set of features. PWs in a query word image are then compared with the PWs in the database to be searched using a three stage matching realizing promising results. Since we work on PWs, the method is not sensitive to perfect character segmentation making it an attractive choice for Urdu text where an ideal word or character segmentation is not achievable. In the present system, each PW of the query word is compared with each PW in the collection of documents to be searched. This may be replaced by first performing a clustering of PWs on the reference data set and then comparing each PW of the query word to a class of PWs instead of each and every PW. In addition, for practical purposes, the information to be searched is not always available in the form of an image. It would therefore be good idea to replace query image by an on-screen keyboard and later by a sketch pad where user may just sketch the query word and expect the system to return the relevant documents. These aspects will form the subject of our subsequent research. It is anticipated that our work would serve as an important contribution in realizing the goal of digitized, searchable Urdu libraries. It is also expected that the proposed work will contribute towards the development of an Urdu OCR system as well.

REFERENCES

[1] Jameson, M., "Promises and challenges of digital libraries and document image analysis: a humanist's perspective", 1st Int'l workshop on document image analysis for libraries, DIAL04.

[2] Rothfeder, J. L., Feng, S., and Rath, T. M., " Using corner features correspondance to rank word images by similarity". Conference on Computer vision and pattern recognition, USA, pp. 30-35.

[3] Gatos, B. and Pratikakis, I., "Segmentation free word spotting in historical printed documents". 10th international conference on document analsis and recognition", 2009.

[4] Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., and Perantonis, S. J., "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback". IJDAR, 9: 167-177.

[5] Adamek, T., O'Connor, N. E., and Smeaton, A. F., "Word matching using single closed contours for indexing handwritten historical documents". IJDAR, 9:153-165.

[6] Madhvanath, S. and Govindaraju, V., "The role of holistic paradigms in handwritten word recognition". IEEE transactions on pattern analysis and machine intelligence, 23:149-164.

[7] Li, J., Fan, Wu, Y., and Le, N., "Document image retrieval with local feature sequences". 10th International Conference on Document Analysis and Recognition.

[8] Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., and Popescu, G.V., " A line oriented approach to word spotting in handwritten documents". Pattern Analysis and Applications, 3: 153-168.

[9] Dehghan M. and Faez K., "Handwritten farsi (Arabic) word recognition: a holistic approach using discrete HMM". Pattern Recognition 34: 1057-1065.

[10] Rath, T. M. and Manmatha, R., "Word spotting for historical documents". IJDAR, 9:139-152, 2007.

[11] Reicher, G. M., "Perceptual recognition as a function of meaning fulness of stimulus material". Journal of Experimental Psychology,275-280.

[12] Leydier, Y., LeBourgeois, F., and Emptoz, H., "Textual indexation of ancient documents". Proceedings of the ACM symposium on Document engineering, 111-117.

[13] Moghaddam, R. F. and Cheriet, M., "Application of multilevel classifiers and clustering for automatic word spotting in historical document images". 10th International Conference on Document Analysis and Recognition.

[14] Rodriguez-Serrano, J. A. and Perronnin, F., "Handwritten word image retrieval with synthesized typed queries". 10th International Conference on Document Analysis and Recognition.

[15] Terasawa, K., Imura, H., and Tanaka, Y., "Automatic evaluation framework for word spotting". 10th International Conference on Document Analysis and Recognition.

[16] Khurshid, K., Faure, C., and Vincent, N., A novel approach for word spotting using merge-split edit distance, CAIP, 2009, Germany

[17] M.W.Sagheer and Nicola Nobile, "A novel handwritten word spotting based on connected component analysis". IEEE, international conference on Pattern Recognition.