Word Spotting based Retrieval of Urdu Handwritten Documents

Ali Abidi
National University of Sciences & Technology
Islamabad, Pakistan
abidi@mcs.edu.pk

Akhtar Jamil
Comsats University
Abbotabad, Pakistan
akhjamil@ciit.net.pk

Imran Siddiqi
Bahria University
Islamabad, Pakistan
imran.siddiqi@bahria.edu.pk

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

Abstract

Urdu being one of the most popular languages adopted during different swatches of history has a valuable collection of handwritten scripts in different state libraries of South Asia. Digitizing these collections can serve not only to preserve them but also to make them available to general public. Non existence of an Urdu OCR, however, limits the concept of a digital Urdu library to scanning and manual search of documents only. We present a word spotting based search method for Urdu handwritten text. The text is first segmented into partial words and a set of features is computed from each partial word. The user queries the system using word image. The partial words in the query image are then matched with those in the database and the matched partial words are merged into complete words. The proposed method evaluated on 90 handwritten documents reported encouraging precision and recall rates.

Keywords*-Urdu handwritten text detection; Partial Words; Run length smoothing alogrithm*

1. Introduction

South Asian libraries hold huge collections of valuable handwritten documents. The digitization of these documents can made them accessible to larger audiences through different forms of electronic media. These digital collections however are very large and unstructured and finding specific information of interest in these collections remains a tedious and time consuming task. An obvious solution to the problem is the manual annotation/transcription of these documents which naturally is a heavy job in terms of time, labor and cost. Research in Optical Character Recognition (OCR) and handwriting recognition has greatly contributed to automate this transcription ultimately allowing efficient retrieval of desired information. An attractive alternative to text/handwriting recognition is word spotting where the information searched for is retrieved by matching the shape of the query word(s) with those in the database without any semantic knowledge of what is being queried. Our present work is dedicated to the same idea of information retrieval in Urdu handwritten documents using word spotting. The presented approach accepts a query word image and retrieves all the handwritten documents containing occurrences of the query word.

Digital libraries have made considerable contributions to reinstate the day by day decreasing importance of a conventional library in our daily lives [12]. In case of handwritten text, word spotting has been an attractive choice [6] as commercially developed OCRs are far from achieving good recognition rates on ancient or handwritten documents [15]. State of the art word spotting techniques are divided into two broad categories: image based matching techniques and feature based matching techniques. Image based matching techniques compute distances between words directly on image pixels [8, 16]. Feature based matching techniques, on the other hand; first compute certain features for word images and then match these features with those in the database [5]. Another known and more practical division is to divide the methods into either segmentation based [11] or segmentation free methods [6]. We have also followed the segmentation based approach in which a document image is segmented into smaller units (PWs) which can be recognized independently or when grouped [4].

The main contribution of this paper is the extension of our existing word spotting system for printed documents [1] to handwritten text which naturally is more challenging due to writer dependent variations in the writing styles. The proposed scheme relies on a set of features including scalar as well as vector features that are extracted from each partial word in the writing. During the retrieval phase, a multi-stage matching technique is used to locate all occurrences of the relevant partial words which are then merged into words.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.289

The methodology evaluated on about 90 handwritten documents reported promising precision and recall rates.

The paper is outlined as follows. We first briefly describe the data set used followed by some of the challenges associated with Urdu text. We then discuss the indexing and retrieval mechanism along with the set of features used. This is followed by the detailed experimental results and analysis and finally we give some concluding remarks along with some suggested enhancements to the present system.

2. Data Collection

Unlike text in other languages, Urdu does not have any standard benchmark data sets to the best of authors' knowledge. For this reason we collected about 90 handwritten documents authored by the same number of writers. Each of these writers copied a given text in his/her natural handwriting. These writings were then scanned at 300dpi, 8 bits/pixel. Each document image contains on the average 120 words.

3. Challenges with Urdu Text

Urdu, the official language of Pakistan and a largely spoken language of India, has more than 100 million speakers. With few exceptions, Urdu script closely resembles Arabic and Persian. Urdu is written from right to left and words are formed by combining various combinations of sub words which we call Partial Words (PWs), where a PW is made up of different combinations of basic Urdu characters. Urdu is one of the most difficult and challenging scripts to deal with. One of the most challenging problems is the position dependent appearances of characters within a word which makes the segmentation and recognition of characters very difficult. Few Urdu alphabets and their different shapes are depicted in Figure 1.

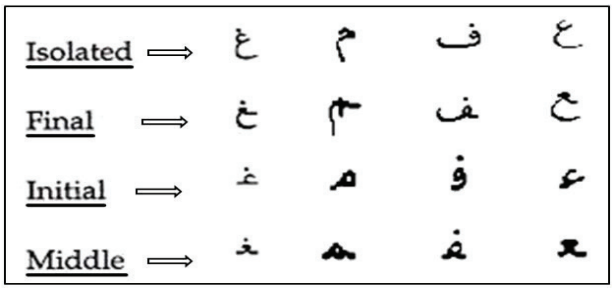

Figure 1. Varying appearances of few Urdu characters.

Another challenge is the non uniform inter and intra word distances. The intra word distances often exceed the inter word distances making it practically impossible to use traditional text segmentation techniques like Run Length Smoothing Algorithm (RLSA) and contour based segmentation methods etc. Figure 2 illustrates a perfect segmentation of Urdu text which is not possible to achieve with the existing state of the art segmentation techniques.

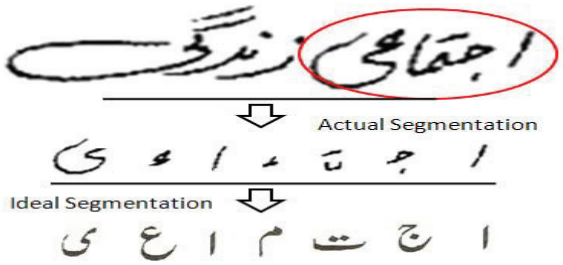

Figure 2. Ideal segmentation is impossible

Other challenges associated with Urdu language include excessive overlapping of adjacent PWs within a word, presence of an excessive number of dots and diacritic marks. For handwritten text, the writer dependent variations in the same words add to these problems.

4. Proposed Methodology

The proposed information retrieval framework is divided into two main parts, indexing and retrieval. Indexing includes segmentation of text into partial words (PWs) and extraction of features from each of the PWs. In retrieval, a multi-stage comparison is carried out between the PWs of the query word image and those in the database to find a match for the query word. Figure 3 shows an overview of our proposed method while each of these two stages is discussed in detail in the following.

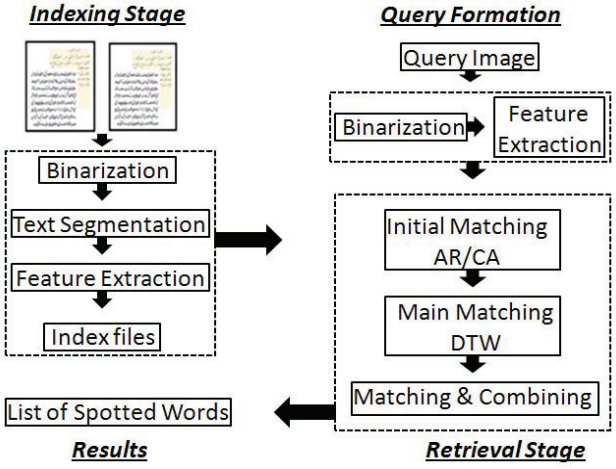

Figure 3. An overview of proposed method

4.1 Indexing

The basic steps of indexing are similar to those we employed in our previous work on printed documents [1]. Since the data set under study comprises contemporary Urdu handwritten text images which are not very noisy, a global thresholding using Otsu's algorithm was carried out

to segment text from background. The next step is to decide the level at which features are to be calculated. Natural choices are word (holistic approach) or character (analytical approach) level features. However, as discussed earlier, the state of the art segmentation schemes cannot be directly applied for word or character segmentation in Urdu text. We therefore decided to work on partial words (PWs). A partial word (PW) is composed of one or more basic Urdu alphabets joined together to form a portion of a word. From the view point of implementation, the partial words are extracted by finding the connected components in the binarized text. Figure 4 shows some PWs extracted from a line of hand written Urdu text.

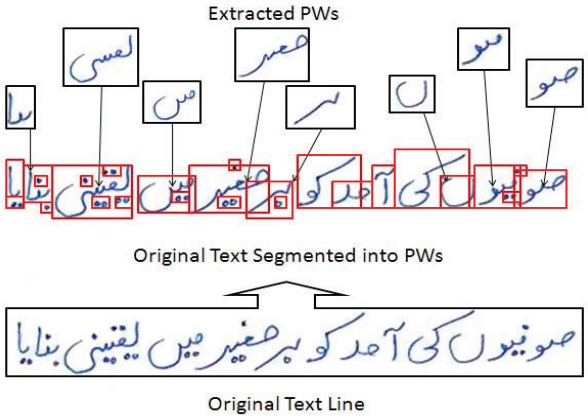

Figure 4. Extraction of PW from handwritten text

For each of the PWs we extract a set of features including scalar features, vertically defined vector features and horizontally defined vector features. The first two of these have already been investigated on printed text [1] while the horizontally defined features have been added to strengthen the feature set. A summary of the proposed features is as follows.

4.1.1. Scalar Features: Scalar features include aspect ratio and convex area of each partial word.

4.1.2 Vertically Defined Features: These features are computed on columns of the PW and include upper profile, lower profile, ink to non-ink transition and vertical projection.

4.1.3 Horizontally Defined Features: The horizontally defined features are calculated on rows of the PW and include; *Right Projection*: the normalized distance of first ink pixel in each row from the right side of the bounding box and *Left Projection*: the normalized distance of last ink pixel in each row from the right side of the bounding box.

These features are aimed at capturing the shape of a PW and have been effectively used on Latin alphabets at word [2] and character [3] levels. All vector feature profiles are obtained from the binary images except the vertical projection which is obtained from gray scale images.

Once the features are extracted, an index file is generated for each document image in the reference base. For each PW in the image, we keep its position within the image, the two scalars and six vertically and horizontally defined vector feature sequences.

4.2 Retrieval

During retrieval, the system is presented with a query word image which is then searched in all the documents in the database. For that, the query word is first segmented into partial words and the set of features discussed earlier is extracted for each PW. A three-stage matching is then carried out to find the instances of the query word in the database. The first two matching stages work on PWs while the last stage takes all the matched PWs and merges them into words based on the correct order of appearance and relative distances between the PWs in a word.

The first matching is done on the basis of scalar features where all the PWs which differ significantly in aspect ratio and convex area from the query PW are eliminated. This filtering not only reduces the search space for next stage but also serve to enhance the performance of the system. In our experiments, 62% of the total PWs are filtered and only 38% are passed for matching in the next stage.

The main matching using the vector features of PWs is carried out using Dynamic Time Warping (DTW) owing to its ability to cater for non-uniform stretch, size and style of two signals to be matched. The (writer-dependent) style variations between the same PWs are catered for by DTW as it allows different styles to be matched by compressing and stretching their profile features with respect to each other. Since the vector features are computed horizontally as well as vertically, DTW is separately applied on each set of features the final distance being computed as an average of the two. All PWs with distances below a predefined threshold to the query PW are accepted.

In the final stage, a sliding window keeps on comparing the spotted PWs along a given line of text and merges them together till the time they appear in the correct order as per the order of PWs in the query word image. Figure 5 shows the spotted PWs for a given query word while Figure 6 shows stepwise merging of PWs into a complete word.

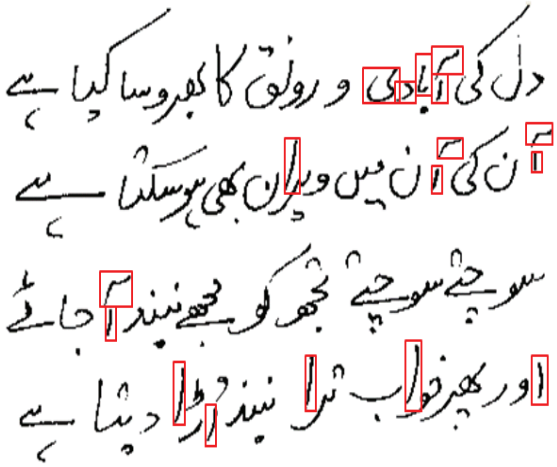

Figure 5. Spotted PWs for a given query word

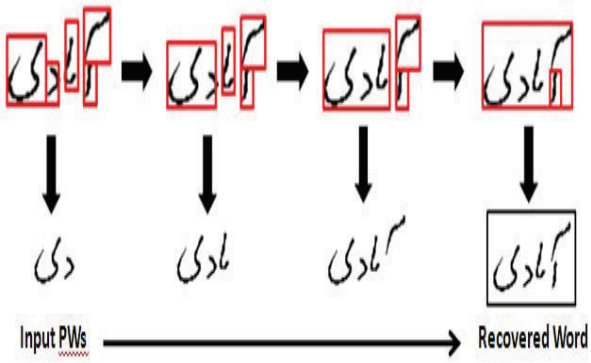

Figure 6: Merging of PWs into word

The output of the final merging is a list of all the documents containing potential instances of the query word. The list is then presented to the user for further inspection.

5. Experimental Evaluation

The proposed system is evaluated on a data set 90 handwritten images containing an average of 120 words per image. For testing, 115 query words having 745 instances in total were selected. These 115 query words are further divided into 387 PWs having 1823 instances in total. Different sets of experiments were conducted to analyze the performance of the proposed features as well as the different matching stages to study their contribution in the overall system performance. These experiments are discussed in the following.

5.1 Indexing Time

Indexing involves computation of features of each PW and storing the feature values in index files. The indexing is carried out offline therefore the indexing time may not be very significant. Nevertheless, Figure 8 shows the indexing time as a function of the number of documents to be indexed.

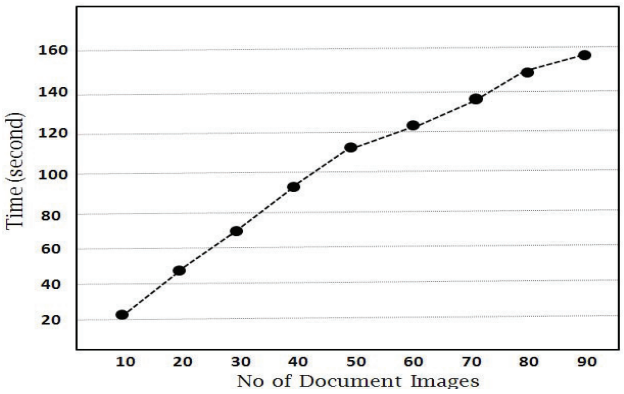

Figure 8. Indexing time vs. number of documents

5.2 Performance of Individual Features

In an attempt to evaluate the discriminative power of the proposed features, we first evaluate each of the features independently and the respective recall rates (on PWs) are indicated in Figure 9. It can be seen that the vertical projection comes out to be the most effective feature correctly matching about 84% of the partial words.

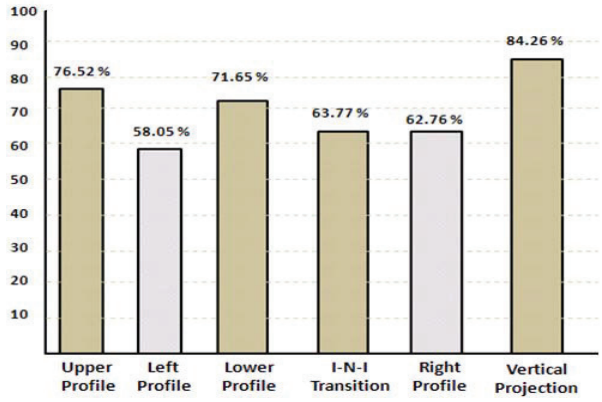

Figure 9. Recall rates- individual features

5.3 Performance of Combined Features & Matching

The overall performance of the system is determined at word level after the three matching stages. We compute the standard precision, recall and F-measure. Figure 10 shows the results of the initial matching stage (PWs), the matching of PWs using DTW and the final matching of complete

words. To show the effectiveness of the initial matching, we have presented the results of DTW matching with and without initial matching as well.

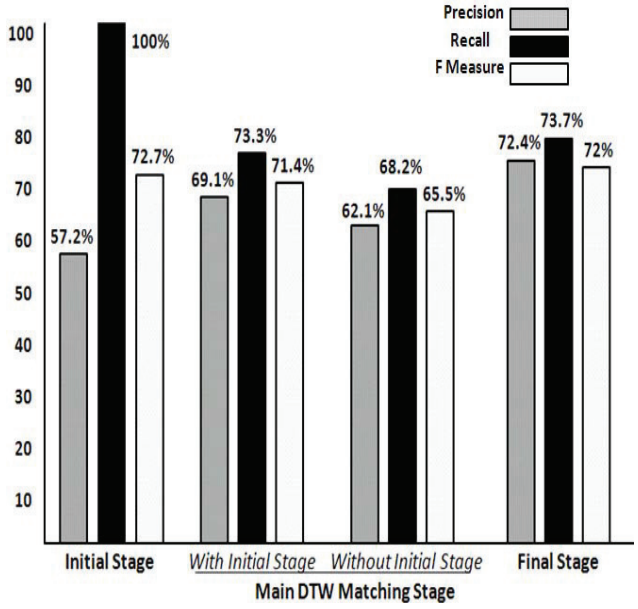

Figure 10. Performance of different matching steps

Figure 11 illustrates the true positives, false negatives and false positives obtained at different matching stages.

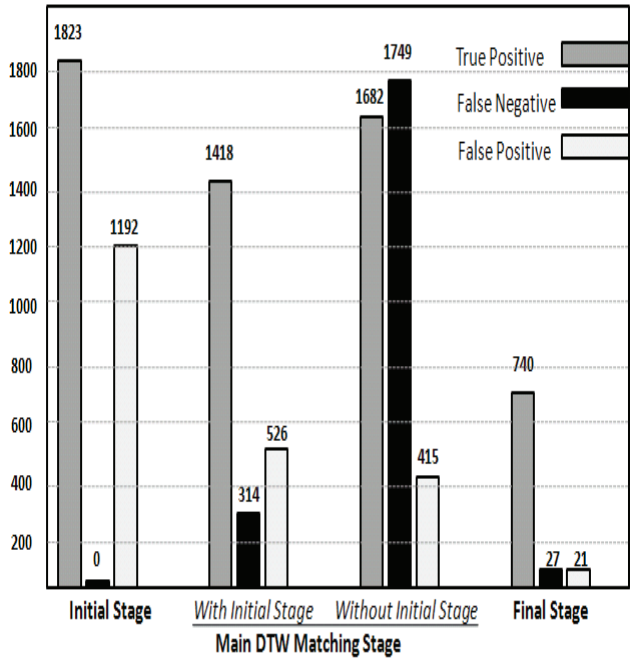

Figure 11. Comparison of different matching stages

5.3 Computational Time

Any retrieval system is graded by the time it takes for finding the desired information. To study the computational time of retrieval, each of the matching stages is individually analyzed. Naturally, this will depend upon the number of PWs in the query image (and of course the size of the database searched which is fixed in our series of experiments). We therefore present to the system query words comprising one to five PWs per word. Figure 12 gives an estimate of the time different stages take while the average retrieval times are listed in Table 1.

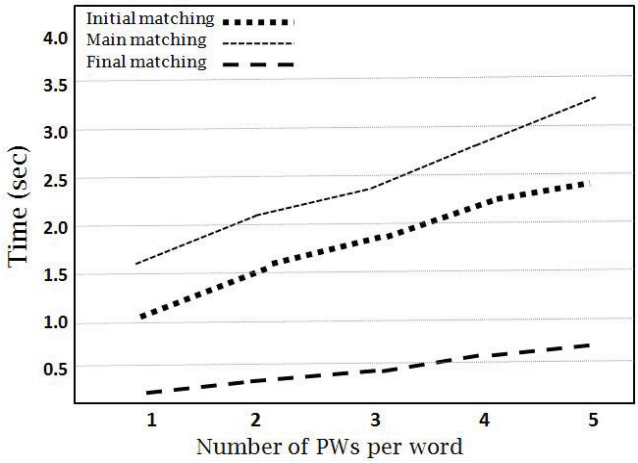

Figure 12. Computational time for matching stages

Table 1. Average retrieval time - PWs per word

No of PWs/ word	Initial matching	Main matching	Final matching	Total time
1	1.025s	1.65s	0.2s	2.875s
2	1.235s	2.05s	0.26s	3.55s
3	1.665s	2.75s	0.35s	4.75s
4	1.985s	3.05s	0.415s	5.45s
5	2.445s	3.3s	0.605s	6.35s

6. Conclusion

We have presented a word spotting based technique fro information retrieval from Urdu handwritten texts. The work is an extension of our previous work on printed Urdu documents. The method relies on extracting a set of features from the partial words and comparing them using two DTW

modules. This work can be complemented by a clustering mechanism where the PWs in the indexed documents are clustered. This will allow matching the query PWs to clusters only and not each and every PW in the database thereby tremendously increasing the computational efficiency of the system. The authors expect that the presented work will contribute towards the practical implementation of searchable Urdu digital libraries.

7. References

[1] Abidi, A., Siddiqi, I., and Khurshid, K., "Towards searchable digital Urdu libraries - a word spotting based retrieval approach", In Proceedings of the 11th International Conference on Document Analysis and Recognition, 2011, pp. 1344-1348.

[2] Rath, T.M., and Manmatha, R.., "Word image matching using dynamic time warping", In Proceedings of the IEEE Computer Vision and Pattern Recognition Conference, 2003, pp. 521.

[3] Khurshid, K., Faure, C., and Vincent, N., "Feature based word spotting in ancient printed documents", In Proceedings of the 8th eition of PRIS in 10th international conference on enterprise information systems, ICEIS, 2008.

[4] Sagheer, M.W., Nobile, N, He, C.L. and Suen C.Y., "A novel handwritten word spotting based on connected component analysis", In Proceedings of the 20th International conference on Pattern Recognition, 2010, pp. 2013-2016.

[5] Manmatha, R., Han, C., and Riseman, E.M., "Word spotting: A new approach to indexing handwriting", Technical report CS-UM-95-105, Computer Science Dept, University of Massachusetts at Amherst, MA, 1995.

[6] Gatos, B. and Pratikakis, I., "Segmentation free word spotting in historical printed documents", In Proceedings of the 10th International conference on document analsis and recognition, 2009, pp. 271-275.

[7] 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(2): 167-177, 2007.

[8] Rath, T. M., Kane, S., Lehman, A., Partridge, E., and Manmatha, R., "Indexing for a digital library of George Washington's manuscripts: a study of word matching techniques", Technical Report, University of Massachusetts, Amherst, 2002.

[9] 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, 2007.

[10] Madhvanath, S. and Govindaraju, V., "The role of holistic paradigms in handwritten word recognition", IEEE transactions on pattern analysis and machine intelligence, 23(2):149-164, 2001.

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

[12] Jameson, M., "Promises and challenges of digital libraries and document image analysis: a humanist's perspective", In Proceedings of the first International workshop on document image analysis for libraries, 2004, pp. 54-61.

[13] Tersawa, K., Imura, H., and Tanaka, Y., "Automatic evaluation framework for word spotting", In Proceedings of the 10th International conference on document analysis and recognition, 2009, pp. 276-280.

[14] Rothfeder, J.L., Feng, S., and Rath, T. M., "Using corner features correspondences to rank word images by similarity", In Proc. of the Workshop on Document Image Analysis and Retrieval (DIAR), 2003.

[15] Leydier, Y., LeBourgeois, F., and Emptoz, H., "Textual indexation of ancient documents", In Proceedings of the ACM symposium on document engineering, 2005, pp. 111-117.

[16] Lewis, J.P., "Fast template matching", Vision interface, pages 120 – 123.Canadian Image Processing and Pattern Recognition Society, 1995, pp. 120-123.

[17] Khurshid, K., Faure, C., and Vincent, N., "A novel approach for word spotting using merge-split edit distance", In Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns, CAIP, 2009, pp. 213-220.

[18] Khurshid, K., Faure, C., and Vincent, N.: Word spotting in historical printed documents using shape and sequence comparisons. Pattern Recognition 45(7):2598-2609, 2012.