Language Independent Keyword Based Information Retrieval System of Handwritten Documents using SVM Classifier and Converting Words into Shapes

Muhammad Rashid Hussain[1], Asif Masood[2], Haris Ahmad Khan[3], Imran Siddiqi[4], Khurram Khurshid[3]

1. National University of Sciences and Technology, Islamabad, Pakistan
2. National University of Sciences and Technology, Islamabad, Pakistan
3. Institute of Space Technology, Islamabad, Pakistan
4. Bahria University, Islamabad, Pakistan

* **Corresponding Author:** E-mail: raashid55@mcs.edu.pk, amasood@mcs.edu.pk

Abstract

This work presents a language independent keyword based document indexing and retrieval system using SVM as classifier. Word spotting presents an attractive alternative to the traditional Optical Character Recognition (OCR) systems where instead of converting the image into text, retrieval is based on matching the images of words using pattern classification techniques. The proposed technique relies on extracting words from images of handwritten documents and converting each word image into a shape represented by its contour. A set of multiple features is then extracted from each word image and instances of same words are grouped into clusters. These clusters are used to train a multi-class SVM which learns different word classes. The documents to be indexed are segmented into words and the closest cluster for each word is determined using the SVM. An index file is maintained for each word containing the word locations within each document. A query word presented to the system is matched with the clusters in the database and the documents containing occurrences of the query word are retrieved. The system realized promising precision and recall rates on the IAM database of handwritten documents.

Key Words: *Word Spotting, Handwritten, Image processing, SVM,Optical Character Recognition (OCR)*

1. Introduction

Information retrieval in textual document images has been an active area of research for the last three decades [1, 2]. Ease in information retrieval enhances the importance and value of digitized corpus manifolds. Optimized and robust systems are being designed to cater for the diverse challenges to make automated information retrieval more efficient and reliable. Handwritten documents are the most challenging of the lot as there exists a vast diversity in handwriting styles across different individuals. Optical Character Recognition (OCR) that yields high recognition rates in case of printed documents is yet to mature for handwritten text. Word spotting is an interesting alternative to traditional OCR systems and has been effectively employed to match word images at various levels to retrieve similar words to a given query image. The basic technique is to extract features from handwritten documents which should be robust to various writing styles and then apply supervised classification by employing classifiers like Neural Networks [3, 4] or Support Vector Machine (SVM) [5, 6] for grouping similar words together.

Word spotting systems are especially effective in the context of digital libraries. The process of digitization of paper content in libraries naturally needs to be complemented with efficient and effective retrieval systems. In situations where the content cannot be converted into text, word spotting could be employed for retrieval purposes. Likewise, the historical manuscripts which suffer from degradations and noise, can also be effectively retrieved against a query word using word spotting techniques. Such documents are hard to segment and recognize and word spotting is the most effective retrieval technique for such documents. Since word spotting treats each word as an image, word spotting systems can also be adapted into more general image retrieval systems.

Various word spotting systems have been proposed in the literature that can be categorized into two major classes based on the level of matching. These are holistic methods and analytical methods. Holistic techniques (sometimes also called segmentation-free techniques) consider word as the basic unit. A number of techniques depending on the type of document/handwriting are employed to extract words from the images. Features are defined at word level and matching is carried out using different similarity measures to retrieve similar words to the given query. Holistic approach is more useful when segmenting the word into characters or other small units is not feasible as is the case in most handwritten documents and has been widely studied in the literature [7-18].

In analytical (segmentation-based)techniques, a word is further segmented into characters or other sub-units like partial words [19] S-characters [20] or graphemes [21] etc. Extracting features at this lower level captures more intrinsic information of the word [22]. However, due to segmentation issues, analytical techniques are more difficult to use for handwritten text as compared to the printed text [23]. Variants of analytical techniques have been employed by a number of researchers [24-30] for word spotting and similar related problems.

Both the techniques, holistic and analytical, have their pros and cons and it is very difficult to compare the two as the comparison becomes subjective to the type of documents and data set. In general, holistic techniques are preferred for highly cursive scripts while analytical techniques are more useful for scripts which offer lesser segmentation challenges.

This paper presents a novel font independent word spotting based indexing and retrieval system for handwritten documents. A set of documents segmented into word images is used to generate a set of word clusters (semi-automated algorithm based on closest match) which serves as a reference base. A multi-class support vector machine is trained on the word clusters using a set of statistical features. The words in the documents to be indexed are matched with the clusters and an index file is maintained for each cluster comprising information on the occurrences of words in different documents. During retrieval, a query word image presented to the system is matched against the stored clusters to retrieve all documents containing instances of the query word.

This paper is organized as follows. In the following section we present in detail the proposed indexing and retrieval methodology. Section 3 describes the implementation methodology for cursive Urdu text to show script independence of the method. Section 4 presents experimental evaluations conducted to validate the proposed methodology along with an analysis of the results realized. Finally, we conclude the paper with some ending remarks.

2. Proposed Methodology

The proposed indexing and retrieval methodology relies on two main phases.In the first phase, words extracted from a set of documents are grouped into clusters using a semi-automated clustering technique. Features extracted from the word clusters are then used to train a multi-class SVM that learns to discriminate between different word classes (Figure 1a).

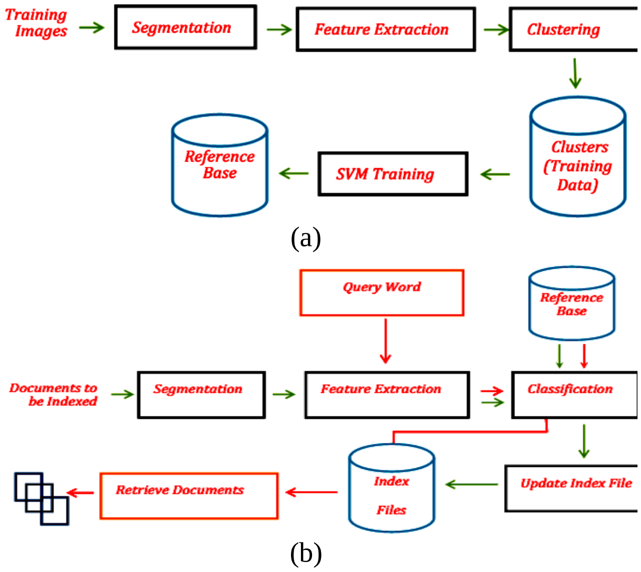


Fig. 1 (a): Generation of word clusters/reference base(b): Indexing and retrieval

The second phase comprises indexing and subsequent retrieval of documents. During indexing, the documents to be indexed are pre-processed and word segmentation is carried out. Each of the extracted words is classified into one of the word classes (clusters) in the reference base and the index file of the respective cluster is updated. During retrieval, a query word image presented to the system is matched with the clusters in the database and the index file of the matched cluster is used to retrieve all the documents containing instances of the query word. The overall indexing and retrieval process is summarized in Figure 1b while each of these steps is detailed in the following sub-sections.

2.1 Segmentation

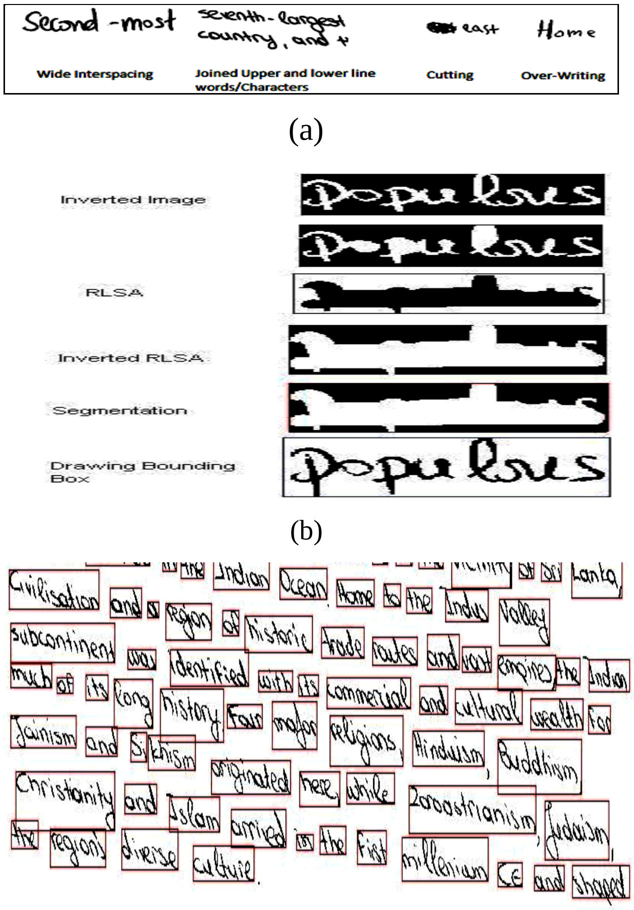
As opposed to printed text, segmentation of words from handwritten text presents a more challenging problem. The main difficulties arise due to varying writing styles of different authors, non-uniform inter and intra word spacing, slanting lines and irregular placement of punctuation marks.

As a first step, the document image needs to be binarized. Since we consider contemporary images of handwriting, a global thresholding with Otsu's algorithm is employed to binarize the image. For extraction of words, we first apply the region filling algorithm on the binarized image of handwriting to fill the loops in characters. The horizontal Run Length Smoothing Algorithm (RLSA) is then applied to merge characters in a word into a single connected component. The RLSA threshold is chosen to be large enough to join characters within a word and small enough not to merge neighbouring words.

Major issues with RLSA based word segmentation include merging of words across different lines and fusion of punctuation marks with the word. These issues are addressed using a set of heuristics on the area and position of the detected bounding boxes. Figure 2a illustrates the steps involved in RLSA based segmentation while segmentation of words from a sample document image is illustrated in Figure 2b. Overall 97% accuracy has been achieved using varying segmentation algorithms.

2.2 Feature Extraction

Process of Transforming Words into Shaped Images. Feature extraction is considered the most critical step in a pattern classification problem that aims to find a characteristic representation of patterns under study (words in our case). In order to cater the varying writing styles of different writers, we propose conversion of words into shapes, eliminating the unnecessary intra-word white spaces and smoothing the word boundaries. Our proposed idea of treating words as shapes arises from the observation that (the space normalized) boundary of a particular word results in more or less the same shape even if it is written in different writing styles. We, therefore,treat each word as a shape and employ shape matching techniques for document retrieval.



(a)

(b)

(c)

Fig. 2 (a):Common problems in handwritten documents (b) Process of RLSA based word segmentation (c): Words segmented from a sample image

Determination of Baseline after Removal of Extra White Spaces. For each extracted word, the empty columns between different components are eliminated and the baseline (row with maximum number of text pixels) of the word is determined.

InterJoining the Components Dynamically. All the components in a word are then joined together by a horizontal line passing through the baseline. Loops and gaps in characters are filled by applying the morphological region filling algorithm to the word image and finally the contour of the word is extracted hence representing each word as a unique shape. Figure 3 illustrates an example of a word extracted from a document and converted into shape.

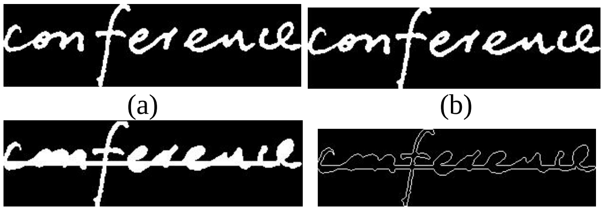

(a) (b)
(c) (d)

Fig. 3 Conversion of word to shape (a): Original image (b): Blank columns removed (c) Closed image with dynamic base line joined (d): Contour of the word shape

Contour information forms the basis of our shaped word. Now instead of applying set of features on each and individual character, a shaped word which is connected in a fixed pattern is used to extract features. Apart from conventional features (used mostly on words / characters and hardly used in shapes) we have also added new features capturing the change in area of a shaped word thus enhancing feature vector for more robust and efficient matching. We have chosen to employ a set of features capturing the shape information like local orientations, curvatures, geometry and other shape descriptors. These features are detailed in the following.

2.2.1 Chain Code Features (f1)

Chain codes have been applied to a number of shape recognition problems with varying degrees of success. In case of document recognition, chain codes have been applied to problems like writer identification and verification [21, 31], character and word recognition [32, 33, 34, 35], classification of writing styles [36] and handwriting based characterization of writer demographics [37]. In our case, we exploit the chain code representation of the contour of the word shape to extract local orientation and curvature information. The contour is represented as a string of Freeman chain codes and a histogram of these codes is computed to be used as a feature. The normalized histogram of chain codes represents the distribution of local orientations where the dominant orientations are reflected as peaks in the histogram.

To capture the local curvature information from the contour, we compute a 2D histogram of chain code pairs where the entry *(i,j)* of the histogram represents the probability of finding the pair *(i,j)* in the contour representation of a word. Extending the same idea further, a 3D histogram of chain code triplets is also computed and is normalized to be used as a feature [21, 31]. Figure 4 illustrates an example word and the corresponding histogram of chain codes (Figure 4a) as well as the 2D histogram of chain code pairs (Figure 4b).

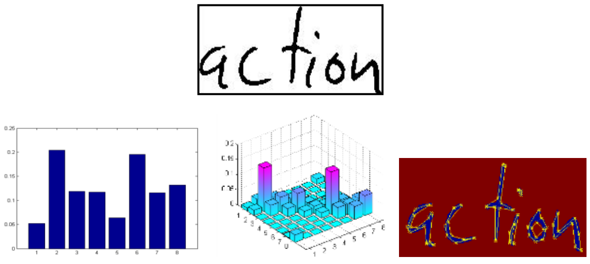

(a) (b) (c)

Fig. 4 An example word and (a): Histogram of chain codes (b): Histogram of chain code pairs (c): Polygonized contours

2.2.2 Polygon Features (f2)

In addition to the chain code representation, the orientation and curvature information of a word is also captured by approximating the word shape by a polygon. Polygonization of word contours not only represents a distant scale of observation but the computed features are also more robust to distortions as compared to the chain code based features. The sequential polygonization algorithm presented in [38] is applied to the contour of a word to represent it by a set of line segments (Figure 4c). The slope of each segment and the curvature (angle) between each pair of neighbouring segments are computed and their distributions are used as features. In addition to these distribution, length weighted distribution of the segment slopes (curvatures) is also calculated where each bin of the histogram is incremented by the length(s) of the segment(s) having a particular slope (angle). These distributions are then normalized to have a sum of 1 and are used as features in characterizing a shaped word [31, 37].

2.2.3 Pixel Density Features (f3)

These features capture the pixel density information in different zones of a word. To compute zone based features (*f3*), a word is divided into four sub-images by placing a 2x2 grid on the image. For each zone, the proportion of pixels with respect to the total number of text pixels in the segmented word is computed.

$$D_i = \frac{N_t}{\sum_{i=1}^{4} N_i}; \; i = 1, 2, 3, 4$$

With N_i being the number of pixels in zone *i*.

2.2.4 Profile Features (f4)

Profile features have been very effectively applied to a number of recognition problems [20, 19]. Among various profile features, upper and lower profiles have been most widely employed. Upper profile is computed by finding, in each column, the distance of the first text pixel from the top of the bounding box of segmented word. In a similar fashion, the lower profile is calculated by finding the distance of the last text pixel from the top of the bounding box. Both the profiles are normalized by dividing them by the height of the segmented word. The dimension of upper and lower profile is the same as the number of columns (width) of the shaped word. Typically, profiles are matched using the well-known Dynamic Time Warping (DTW) method [10]. In our implementation, however, we keep the mean and standard deviation of each profile giving a four dimensional profile feature*f4*.

2.2.5 Projection Features (f5)

Horizontal and vertical projections of a shaped word are computed through summation of total number of text pixel in each row (column) of the shaped word. The projections are then normalized by division with the width (height) of the image. Similar to profile features, we compute the mean and standard deviation of the horizontal and vertical projections and employ them as features.

2.2.6 Zone based Orientation Features (f6)

Similar to the orientation features computed from the polygonzied contours of a word, we also compute the local orientation information in small zones of the word. The word image is converted into skeleton and is divided into 9 (3 × 3) equal sized windows. Features are then extracted from each zone of the word. These features include the number of horizontal, vertical, left diagonal and right diagonal lines. In addition, we also compute the normalized length of all horizontal, vertical, left diagonal and right diagonal lines and the normalized area of the word skeleton.

2.2.7 Shape Descriptors (f7)

These features capture the shape properties of the word image and include the well-known Hu moments, Euler number, the ratio of total number of ink pixels to the total number of pixels and eccentricity of the shape.

2.2.8 Delta Features (f8)

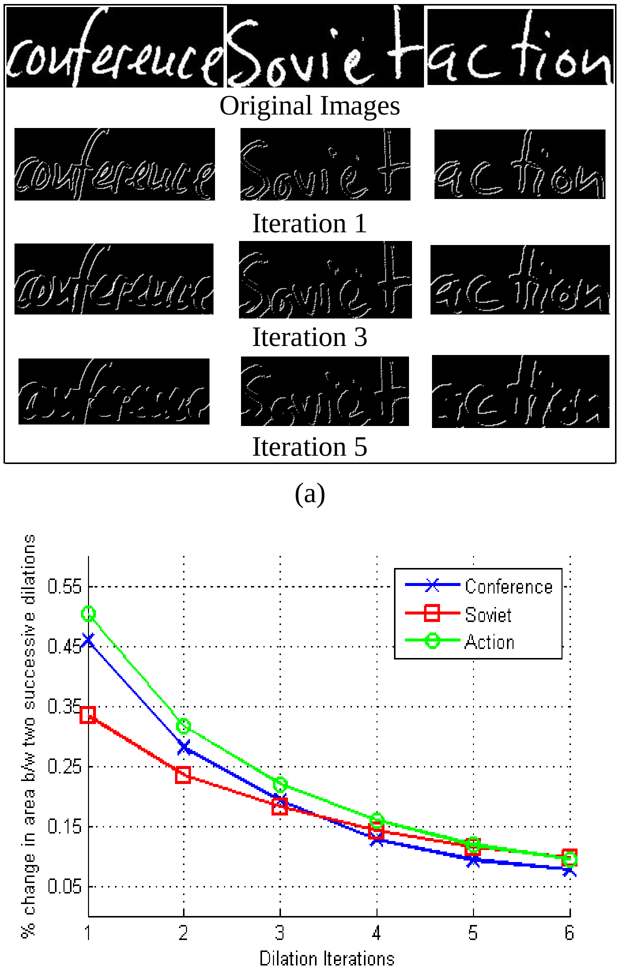
Apart of conventional features, new set of features has also been introduced. These features are aimed at capturing the information on loops, holes and overall structure of the word. Morphological closing is applied on the image to close holes and gaps and the (normalized) difference in area between the original and the closed word image is computed. This value is relatively high for words with loops and holes and serves to discriminate them from other words. In addition, to capture the general structure of the word, we iteratively apply morphological dilation on the word image with 4 different structuring elements (shown in Table 1) and the percentage change in area of the word between two successive dilations is used as feature.

Table 1 Structuring elements used for dilation

$SE1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$SE2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$
$SE3 = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$	$SE4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

The difference in area between two successive iterations of dilation (with one of the structuring elements) on three sample words is illustrated in Figure 5a while the respective percentage area changes for the three words are presented in Figure 5b. In all cases, the increase in area is relatively higher for the initial iterations and stabilizes gradually as the number of iterations increases.

Depending upon the characters within a word, the change in area of the word is different and can be exploited to characterize the word. In our implementation, we carried out 6 iterations of dilation giving a total of 24 features for the four structuring elements.



(a)

(b)

Fig. 5 (a): Change in area of words after dilation with one of the structuring elements (b): Percentage change in area of three words as a function of number of dilation iterations

2.3 Clustering of Words

Prior to indexing of documents, clusters of words need to be generated where each cluster is intended to contain multiple instances of the same word to capture the writer-dependent variations within a word. This clustering is carried out offline, serves the subsequent steps of indexing and retrieval and can be manual, automatic or semi-automatic. In our implementation, we have employed two varying techniques, one is dynamic and second one is semi-automated. For Urdu Language which is extremely cursive, dynamic clustering has been applied and for English text we have experimented using semi-automated technique. In both techniques, clusters generated are manually corrected prior to indexing, if deemed necessary.

Table 2 summarizes all the features considered in our study along with the dimension of each.

Table 2 Summary of features

Feature	Description	Computed From	Dimension
f1	Chain Code based features	Shaped Word	615
f2	Polygon Features	Shaped Word	42
f3	Pixel Density Features	Shaped Word	4
f4	Profile Features	Original Word	4
f5	Projection Features	Original Word	4
f6	Zone based Orientation Features	Skeleton of Word	55
f7	Shape Features	Shaped Word	10
f8	Delta Features	Original Word	25
		Total	**759**

Clustering in English Text. To generate word clusters, we take samples of 20 document images, segment each image into words and represent each word by a feature vector as discussed in the previous section. We have employed a sequential clustering algorithm [31, 39] which does not require a priori the number of clusters. We start by randomly picking a word and assuming it to be the mean (representative) of the first cluster. For each subsequent word, we compute its distance with the centre of each cluster and chose the nearest cluster as a potential candidate. If the distance of the word in question to the nearest cluster is below an empirically determined threshold, the word is assigned to this cluster and the cluster mean is updated. In case the distance does not fall below the predefined threshold, a new cluster is created with the word in question as the mean of the newly generated cluster. This process is repeated until all the words have been assigned to a cluster.

The most significant of short coming of the used clustering algorithm is that the generated clusters are sensitive to the order in which the words are presented to the algorithm. However, it should be noted that the objective of clustering is to generate an approximate set of word classes which are manually corrected prior to indexing. Hence, the performance of the overall system is not sensitive to this clustering step.

Executing the mentioned clustering algorithm on the sample images, we get a total of 136 clusters. These clusters, naturally, contain some errors which are corrected manually. Similarly, clusters with less than 5 elements are removed. After refinement, we come up with 88 clusters containing a total of 941 words. These clusters are then employed to train the SVM based classifier which is used to index the given set of documents. An example from a sample cluster is given below. It contains word 'bordered' in varying styles and written by different writers. Index file is also maintained alongwith clusters which contains information of location of this word in different documents.

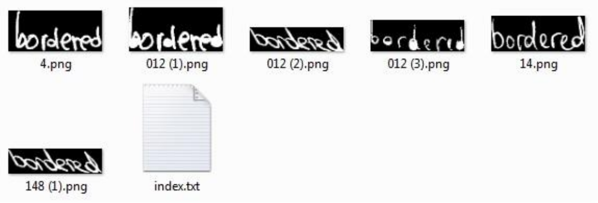


Fig. 6 A sample cluster of words with index file

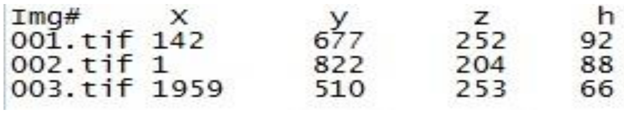
Words without Clusters. It may be noted that it is not possible to have a cluster for each and every new / incoming word already existent in the database; so all the orphan words which donot belong to any cluster are placed in a separate cluster and its sorting will be done subsequently; wherein separate clusters will be generated dynamically.

2.4 Support Vector Machine Training

Once the clusters of words are generated, the features extracted from these words are used to train a multi-class support vector machine (SVM) to learn to discriminate between different word classes. SVM is innovatively and efficiently trained using one-against-all implementation using the radial basis kernel function while the parameters of SVM (C and gamma) are empirically chosen through cross validation. Features extracted from a word are fed to the trained SVM which assigns a probability of classification to each of the classes (clusters of words). The word is assigned to the cluster for which SVM reports the maximum probability. In case the probability of assignment is less than a pre-defined threshold, the word does not belong to any of the clusters in the database and is discarded. If a proper match is not found, information message is given to user that the intended word is not contained in this document.

2.5 Indexing and Retrieval

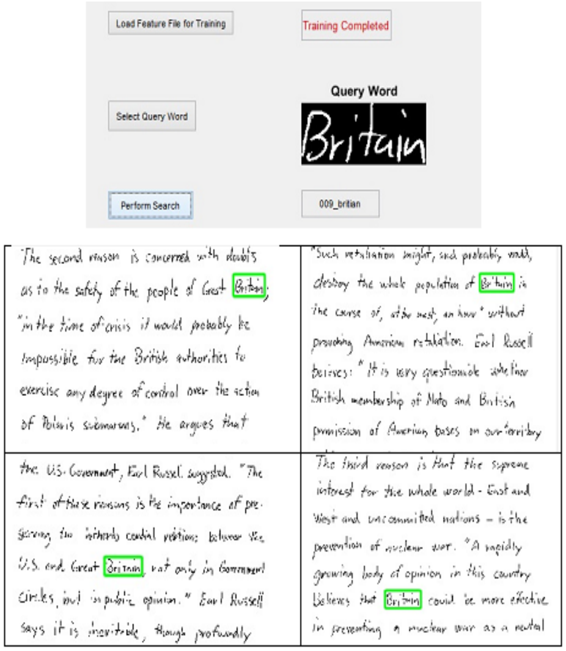
Indexing: Once the SVM is trained to learn to discriminate between different word classes, we proceed to indexing and retrieval steps. During indexing, a set of documents presented to the system is pre-processed, words are segmented and features are extracted from each word as discussed earlier. The features extracted from a word are fed to the trained SVM which outputs the probability scores of the word belonging to each of the clusters (classes). The word is attributed to the class for which the SVM reports the maximum score. Applying a threshold on the confidence score of the SVM to reject the words which are not likely to belong to any of the word cluster adds flexibility to existing system. In case the confidence score of the nearest word cluster is above the threshold, we assign the word to the respective cluster and update the index file of the cluster. The index file contains information on the document image identification and the coordinates of the word within that image. The index file is updated every time a new word is added to the cluster. The index file (Figure 7) of a cluster allows keeping track of all instances of the respective word within the indexed set of documents.



```
Img#      x          y        z       h
001.tif 142        677      252      92
002.tif 1          822      204      88
003.tif 1959       510      253      66
```

Fig. 7 A sample index file for a cluster

Retrieval: During retrieval, a query word image is presented to the system and the idea is to retrieve all documents containing occurrences of the provided word. Features are extracted from the query word image and the SVM is used to find the most probable word cluster that matches the query word. Once the cluster is identified, the index file of the respective cluster is parsed to retrieve all documents containing instances of the query word. Each document containing the query word is displayed to the user with the queried word highlighted on the document. Figure 8 illustrates a retrieval session with the system where the query word 'Britain' is provided to the system and the documents containing instances of the word are retrieved and presented to the user.



"The second reason is concerned with doubts
as to the safety of the people of Great Britain,
"in the time of crisis it would probably be
impossibile for the British authorities to
exercise any degree of control over the action
of Polaris submarines." He argues that

"Such retaliation might, and probably would,
destroy the whole population of Britain in
the course of, at the most, an hour" without
provoking American retaliation. Earl Russell
believes: "It is very questionable whether
British membership of Nato and British
permission of American bases on our territory

the U.S. Government, Earl Russel suggested. "The
first of these reasons is the importance of pre-
serving the hitherto cordial relations between the
U.S. and Great Britain, not only in Government
circles, but in public opinion." Earl Russell
says it is inevitable, though profoundly

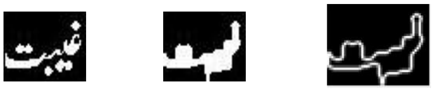
The third reason is that the supreme
interest for the whole world - East and
West and uncommitted nations - is the
prevention of nuclear war. "A rapidly
growing body of opinion in this country
believes that Britain could be more effective
in preventing a nuclear war as a neutral

Fig. 8 A retrieval session with the system

3. Application of Proposed Methodology on Urdu Text

Existing work on Urdu word spotting [19] (largest spoken language in South Asia) suffers from two major limitations. Firstly, indexing is performed using a single file which contains features and relevant locations of all the partial words called ligatures in complete documents dataset. Secondly, retrieval process is very cumbersome and computationally inefficient as the query word is matched with each entry in the dataset to get the required information [19]. This problem gets pronounced for large datasets.

The proposed methodology addresses both of these issues. The only difference with the technique presented for English text is that for Urdu text we need to work on parts of words called ligatures rather than complete words (Figure 9). During retrieval, the query word is divided into ligatures and each ligature is matched against the clusters in the database. The matched ligatures are then re-grouped into words using morphological operations. Finally, as a validation step, both query and retrieved words are converted into shapes and are matched to minimize the false positives. The final outcome is then presented to the user containing occurrences of the query word in complete dataset. Figure 10 shows a retrieval session with the system.



(a) (b) (c)

Fig. 9 An Urdu ligature converted to shape

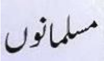
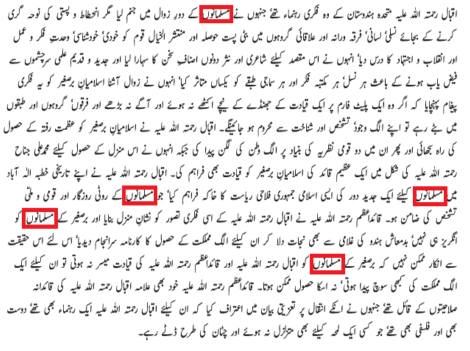
Query Word	Matching
	 اقبال رحمتہ اللہ علیہ متحدہ ہندوستان کے وہ فکری رہنماء تھے' جنہوں نے مسلمانوں کے دورِ زوال میں جنم لیا مگر انحطاط و پستی کی نوحہ گری کرنے کے بجائے نسلی' لسانی' فرقہ ورانہ اور علاقائی گروہوں میں بٹی پست حوصلہ اور منتشر الخیال قوم کو خودی' خودشناسی' وحدتِ فکر و عمل اور انقلاب و اجتہاد کا درس دیا' انہوں نے اس مقصد کیلئے شاعری اور نثر دونوں اصنافِ سخن کا سہارا لیا اور جدید و قدیم علمی سرچشموں سے فیض یاب ہونے کے باعث ہر نسل' ہر مکتبہ فکر اور ہر سماجی طبقے کو یکساں متاثر کیا' انہوں نے زوال آشنا اسلامیانِ برصغیر کو یہ فکری پیغام پہنچایا کہ اگر وہ ایک پلیٹ فارم پر ایک قیادت کے جھنڈے کے نیچے اکٹھے نہ ہوئے اور آگے نہ بڑھے اور فرقوں' گروہوں اور طبقوں میں بٹے رہے تو اپنے الگ وجود' تشخص اور شناخت سے محروم ہو جائینگے۔ اقبال رحمتہ اللہ علیہ نے اسلامیانِ برصغیر کو عظمت رفتہ کے حصول کی راہ سجھائی اور پھر ان میں دو قومی نظریہ کی بنیاد پر الگ وطن کی لگن پیدا کی جبکہ انہوں نے اس منزل کے حصول کیلئے محمدعلی جناح رحمتہ اللہ علیہ کی شکل میں ایک عظیم قائد کی اسلامیانِ برصغیر کو قیادت بھی فراہم کی۔ اقبال رحمتہ اللہ علیہ نے اپنے تاریخی خطبہ الہ آباد میں مسلمانوں کیلئے ایک جدید دور کی ایسی اسلامی جمہوری فلاحی ریاست کا خاکہ فراہم کیا' جو مسلمانوں کے روٹی روزگار اور قومی و ملی تشخص کی ضامن ہو۔ قائداعظم رحمتہ اللہ علیہ نے اقبال رحمتہ اللہ علیہ کے اسی فکری تصور کو نشانِ منزل بنایا اور برصغیر کے مسلمانوں کو انگریز ہی نہیں' بدمعاش ہندو کی غلامی سے بھی نجات دلا کر ان کیلئے الگ مملکت کے حصول کا کارنامہ سرانجام دیدیا' اس لئے اس حقیقت سے انکار ممکن نہیں کہ برصغیر کے مسلمانوں کو اقبال رحمتہ اللہ علیہ اور قائداعظم رحمتہ اللہ علیہ کی قیادت میسر نہ ہوتی تو ان کیلئے ایک الگ مملکت کی کبھی سوچ پیدا ہوتی' نہ اسکا حصول ممکن ہوتا۔ قائداعظم رحمتہ اللہ علیہ خود بھی علامہ اقبال رحمتہ اللہ علیہ کی قائدانہ صلاحیتوں کے قائل تھے' جنہوں نے انکے انتقال پر تعزیتی بیان میں اعتراف کیا کہ ان کیلئے اقبال رحمتہ اللہ علیہ ایک رہنماء بھی تھے' دوست بھی اور فلسفی بھی تھے' جو کسی ایک لمحہ کیلئے بھی متزلزل نہ ہوئے اور چٹان کی طرح ڈٹے رہے۔

Fig. 10 A retrieval session with the system

4. Experiments and Results

The experimental study of the proposed system was carried out on the IAM handwriting database [40]. For indexing we employed a set of 50 document images containing approximately 1000 words. The words in these documents were segmented and compared with the clusters in the database generating index files for each of the clusters. For retrieval, a total of 150 query keywords were presented to the system and the performance was recorded in terms of precision, recall and F-measure. It should be noted that only those words were presented as query for which the corresponding clusters exist in the database.

In order to study the effectiveness of the proposed conversion of word image into a contoured shape, we carried out the evaluations with and without conversion. We also tested the system by providing images containing segmentation errors. The results of these evaluations are presented in Table 3. Using our proposed methodology of transforming words into shapes and employing additional features; we achieved an overall precision of 84% and recall of 89%. Using the original word image to extract features, the precision and recall read 72% and 76% respectively. The results clearly show the effectiveness of the proposed conversion of word into a contoured shape before extraction of features. It is also worth mentioning that a major proportion of errors resulted due to segmentation errors. We achieved more than 95% results once system was tested on good quality images where segmentation errors were negligible. Likewise, high values of precision and recall for Urdu text as summarized in Table 4 demonstrate the language independence of the proposed technique.

Table 3 Retrieval Results on Hand Written English Text

	Feature	Precision	Recall	F-Measure
Proposed Technique	With Shaped Words	0.84	0.89	0.86
	Without Shaped Words	0.72	0.76	0.74

Table 4 Results on Urdu Text (a) Without Shaped Feature (b) With shaped Feature

Query Instances	True Positives	False Positives	False Negatives	Recall	Precision
172	**164**	**18**	**8**	**95.34 %**	**90.10%**

(a)

Query Instances	True Positives	False Positives	False Negatives	Recall	Precision
172	**164**	**6**	**8**	**95.34 %**	**96.47%**

(b)

Performance Sensitivity: In order to study the performance sensitivity of the system to various parameters, we carried out a number of experiments. Once such experiment was to study the system performance evolution as a function of the number of clusters in the database. The number of query words in each case was fixed to 35. The results of these evaluations are summarized in Figure 11. Naturally, the precision and recall are higher for a smaller number of clusters and show a gradual but not dramatic decrease as the number of clusters increases. Table 5 shows the true positives and false positives for different values of threshold and the corresponding ROC curve is illustrated in Figure 12.

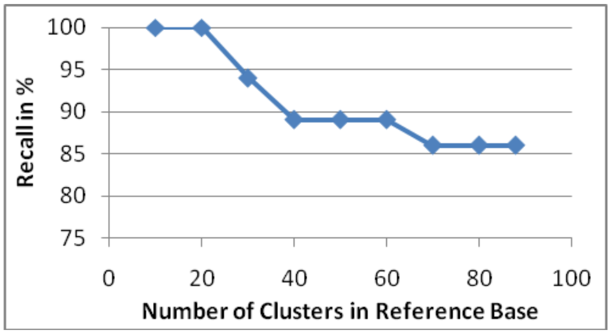


Fig. 11 Recall rates as a function of number of clusters in the reference base

Table 5 ROC Table

Threshold	TP	FP	FN	FP	TP
0.99	147	0	428	0.000	0.256
0.95	300	0	275	0.000	0.522
0.9	479	0	96	0.000	0.833
0.85	502	1	72	0.001	0.873
0.75	508	3	64	0.002	0.883
0.65	515	6	54	0.005	0.896
0.55	525	8	42	0.006	0.913
0.45	525	9	41	0.007	0.913
0.35	525	11	39	0.008	0.913
0.25	525	17	33	0.013	0.913
0.15	531	24	20	0.018	0.923
0.1	532	30	13	0.023	0.925
0.05	538	32	5	0.024	0.936

0	539	36	0	0.027	0.937

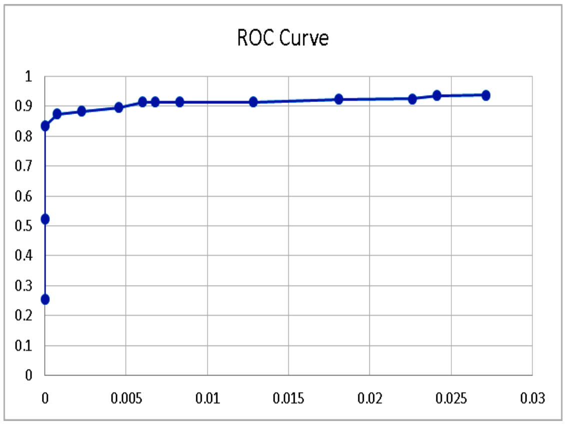
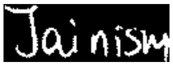


Fig. 12 ROC Curve showing comparison of FP and TP

Table 6 Retrieval times of sample query words using HP core i3 2.40 GHz machine with 8 GB RAM

Query Word	***Time (secs)***
	***1.96***
	3.92
	1.39
	1.43
	3.57
	1.1

To provide an estimate of the processing time, we summarize the retrieval times for few of the query words in the database in Table 6. We also present a comparative analysis of the performance of the proposed system with few well-known word spotting systems reported in the literature. It should be noted that an objective comparison of different systems is difficult due to different databases, different number of samples and different number of query words considered in these studies. Nevertheless, we attempt to summarize the characteristics and performances of some popular word spotting systems (Table 7) to provide readers with a general idea of the performance of these systems. The proposed system realizes precision and recall of 84% and 89% respectively and these values are superior to those reported in the other studies.

Table 7 Comparison of well-known word spotting systems

Study	**Type of Approach**	**Features**	**Classifier**	**DB**	**Results**
Safwan et al. [42]	Line based	Gradient features	Hidden Markov Model	IAM (English), AMA (Arabic) and LAW (Devanagari)	Average Precision = 60%
Frinken et al. [43]	Line based	CTC Token Passing algorithm	Neural Networks	IAM, GW and PARZIVAL	PrecisionIAM=76% GW=71% PARZIVAL=92%
Serrano et al. [44]	Word level segment-ation	Means of the Gaussians	Hidden Markov Model	GW and IFN/ENIT	Average Precision = 91%
Fischer et al [45]	Line based	Geometrical features	Hidden Markov Model	IAM, GW	Average Precision IAM= 55% GW = 74%
Kumar et al. [46]	Line based	Gradient, Structural Concavity & Intensity features	Bayesian logistic regression classifier	IAM (English), AMA (Arabic) and LAW (Devanagari)	Average Precision IAM = 49% AMA = 54% LAW = 51%
Ranjan et al [47]	Word level segment-ation	Profile features	Support Vector Machine	Custom English documents	Average Precision = 81%
Almazan et al. [48]	Word level segment-ation	SIFT	Support Vector Machine	IAM, GW and IIIT5K	Average Precision IAM = 55.73% GW = 92.90% IIIT5K = 72.28%
Proposed method	Word Level segment-ation	Shape Descript-ors	Support Vector Machine	IAM	Precision= 84% Recall = 89%

We also compared our performance with the word spotting system for Urdu text presented in [19]. It should be noted that the system in [19] creates a single index file for the complete database and does not involve any clustering. In other words, every query word is matched with each and every word in the database hence making the retrieval very slow. For large databases, as the number of documents grows, such systems cannot be used for practical purposes. Our proposed system, on the other hand, relies on clusters of words (ligatures). Hence, each query word (ligature) is only matched with the clusters in the database making the retrieval much more efficient.

We also studied the performance of a well-known commercial OCR system ABBYY FineReader [41] which provides high recognition rate rates on printed documents. A total of 50 documents from IAM database containing 14,716 words were fed to the recognition engine. Among these, only 267 words were correctly recognized reflecting the fact that the OCR technology needs to be matured significantly for cursive text recognition. Figure13 illustrates few sample handwritten images and the respective OCR outputs.

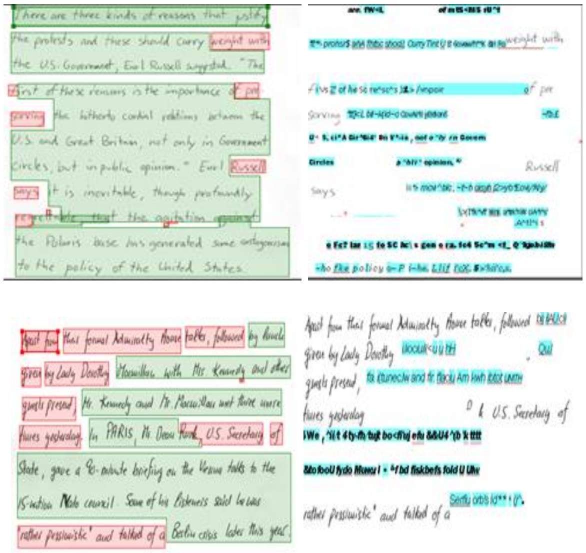


Fig. 13 Recognition results using OCR

3. Conclusion

Our innovative technique of transforming words into shapes and enhancement in feature vector allowed an effective word spotting based document indexing and retrieval system. Segmented words are grouped into clusters based on set of features. These clusters are used to train one-to-all multi-class support vector machine (SVM). The documents to be indexed are segmented into words and the nearest match cluster for each word is determined using the SVM. The index file of the respective cluster is updated to keep information about the document and the location of the word in the document. During retrieval, a query word presented to the system is matched against the clusters in the database and the index file of the matched cluster is used to retrieve the documents containing instances of the query word. The proposed scheme evaluated on the handwritten images of the IAM database realized promising precision and recall rates. The system was also evaluated on printed Urdu documents and the realized results demonstrate the script independence of the proposed technique.

Our future work on this subject will focus on exploring novel features as well as combination of multiple classifiers to further enhance the retrieval performance. A feature selection scheme to study the most effective set of features for this problem is also intended to be carried out.

References

[1] Vinciarelli, A.: A survey on off-line cursive word recognition. Pattern recognition 35(7), 1433–1446 (2002)

[2] Plamondon, R., Srihari, S.N.: Online and off-line handwriting recognition: a comprehensive survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(1), 63–84 (2000)

[3] Frinken, V., Fischer, A., Manmatha, R., Bunke, H.: A novel word spotting method based on recurrent neural networks. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(2), 211–224 (2012)

[4] Liu, Y., Xu, M., Cai, L.: Improved keyword spotting system by optimizing posterior confidence measure vector using feed-forward neural network. In: Neural Networks (IJCNN), 2014 International Joint Conference On, pp. 2036–2041 (2014).

[5] Tarafdar, A., Pal, U., Roy, P.P., Ragot, N., Ramel, J.-Y.: A two-stage approach for word spotting in graphical documents. In: 12th International Conference On Document Analysis and Recognition (ICDAR), 2013 pp. 319–323 (2013).

[6] Impedovo, S., Mangini, F.M., Pirlo, G., Barbuzzi, D., Impedovo, D.: Voronoi

tessellation for effective and efficient handwritten digit classification. In: Document Analysis and Recognition (ICDAR), 2013 12th International Conference On, pp. 435–439 (2013).

[7] Li, J., Fan, Z.-G., Wu, Y., Le, N.: Document image retrieval with local feature sequences. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 346–350 (2009).

[8] Andreev, A., Kirov, N.: Word image matching based on hausdorff distances. In: Proc. 10th International Conference on Document Analysis and Recognition, pp. 396–400 (2009).

[9] Rothfeder, J.L., Feng, S., Rath, T.M.: Using corner feature correspondences to rank word images by similarity. In: Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference On, vol. 3, pp. 30–30 (2003).

[10] Adamek, T., O'Connor, N.E., Smeaton, A.F.: Word matching using single closed contours for indexing handwritten historical documents. International Journal of Document Analysis and Recognition (IJDAR) 9(2-4), 153–165 (2007)

[11] Marinai, S., Faini, S., Marino, E., Soda, G.: Efficient word retrieval by means of som clustering and pca. In: Document Analysis Systems VII, pp. 336–347. Springer, (2006)

[12] Gatos, B., Pratikakis, I.: Segmentation-free word spotting in historical printed documents. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 271–275 (2009).

[13] Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal of Document Analysis and Recognition (IJDAR) 9(2-4), 139–152 (2007)

[14] Zagoris, K., Papamarkos, N., Chamzas, C.: Web document image retrieval system based on word spotting. In: Image Processing, 2006 IEEE International Conference On, pp. 477–480 (2006).

[15] Rusi˜nol, M., Llad´os, J.: Word and symbol spotting using spatial organization of local descriptors. In: Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop On, pp. 489–496 (2008).

[16] Bai, S., Li, L., Tan, C.L.: Keyword spotting in document images through word shape coding. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 331–335 (2009).

[17] Bertolami, R., Gutmann, C., Bunke, H., Spitz, A.L.: Shape code based lexicon reduction for offline handwritten word recognition. In: Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop On, pp. 158–163 (2008).

[18] Kluzner, V., Tzadok, A., Shimony, Y., Walach, E., Antonacopoulos, A.: Word-based adaptive ocr for historical books. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 501–505 (2009).

[19] Abidi, A., Siddiqi, I., Khurshid, K.: Towards searchable digital urdu libraries-a word spotting based retrieval approach. In: Document Analysis and Recognition (ICDAR), 2011 International Conference On, pp. 1344–1348 (2011).

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

[21] Siddiqi, I., Vincent, N.: A set of chain code based features for writer recognition. In: In Proc. of 10th International Conference on Document Analysis and Recognition, pp. 981–985 (2009).

[22] Khurshid, K., Faure, C., Vincent, N.: Feature-based word spotting in ancient printed documents. In: PRIS, pp. 193–198 (2008)

[23] Lu, Y., Shridhar, M.: Character segmentation in handwritten words—an overview. Pattern recognition 29(1), 77–96 (1996)

[24] Terasawa, K., Imura, H., Tanaka, Y.: Automatic evaluation framework for word spotting. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 276–280 (2009).

[25] Vamvakas, G., Gatos, B., Stamatopoulos, N., Perantonis, S.J.: A complete optical character recognition methodology for historical documents. In: Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop On, pp. 525–532 (2008).

[26] Moghaddam, R.F., Cheriet, M.: Application of multi-level classifiers and clustering for automatic word spotting in historical document images. In: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference On, pp. 511–515 (2009).

[27] Leydier, Y., LeBourgeois, F., Emptoz, H.: Textual indexation of ancient documents. In: Proceedings of the 2005 ACM Symposium on Document Engineering, pp. 111–117 (2005).

[28] Frinken, V., Fischer, A., Bunke, H., Manmatha, R.: Adapting blstm neural network based keyword spotting trained on modern data to historical documents. In: Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference On, pp. 352–357 (2010).

[29] Khurshid, K., Faure, C., Vincent, N.: A novel approach for word spotting using merge-split edit distance. In: Computer Analysis of Images and Patterns, pp. 213–220 (2009).

[30] Fischer, A., Keller, A., Frinken, V., Bunke, H.: Hmm-based word spotting in handwritten documents using subword models. In: Pattern Recognition (icpr), 2010 20th International Conference On, pp. 3416–3419 (2010).

[31] Siddiqi, I., Vincent, N.: Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recognition 43(11), 3853–3865 (2010)

[32] Nakano, H.Y.Y.: Cursive handwritten word recognition using multiple segmentation determined by contour analysis. IEICE Transactions on Information and Systems E79-D(5), 464–470 (1996)

[33] Kimura, F., Kayahara, N., Miyake, Y., Shridhar, M.: Machine and human recognition of segmented characters from handwritten words. In: In Proc. of the 4th International Conference on Document Analysis and Recognition, pp. 866–869 (1997)

[34] Blumenstein, M., Verma, B., Basli, H.: A novel feature extraction technique for the recognition of segmented handwritten characters. In: In Proc. of the Seventh International Conference on Document Analysis and Recognition, pp. 137–141 (2003)

[35] Blumenstein, M., Liu, X.Y., Verma, B.: An investigation of the modified direction feature for cursive character recognition. Pattern Recognition 40(2), 376–388 (2007)

[36] M.E.Dehkordi, N.Sherkat, T.Allen: Handwriting style classification. International Journal of Document Analysis and Recognition 6, 55–74 (2003)

[37] Siddiqi, I., Djeddi, C., Raza, A., Souici-meslati, L.: Automatic analysis of handwriting for gender classification. Pattern Analysis and Applications (2014)

[38] Wall, K., Danielsson, P.-E.: A fast sequential method for polygonal approximation of digitized curves. Computer Vision, Graphics, and Image Processing 28(3), 220–227 (1984)

[39] Bensefia, A., Paquet, T., Heutte, L.: A writer identification and verification system. Pattern Recognition Letters 26(13), 2080–2092 (2005)

[40] Marti, U.-V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition 5(1), 39–46 (2002)

[41] AbbyyFinereader, Online: http://www.abbyy.com/finereader/

[42] Wshah, Safwan, Gaurav Kumar, and VenuGovindaraju. "Script independent word spotting in offline handwritten documents based on hidden markov models." Frontiers in Handwriting Recognition (ICFHR), (2012).

[43] Frinken, Volkmar, et al. "A novel word spotting method based on recurrent neural networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, 34.2 (2012): 211-224.

[44] Rodríguez-Serrano, José A., and FlorentPerronnin. "A model-based sequence similarity with application to handwritten word spotting." IEEE Transactions on Pattern Analysis and Machine Intelligence 34.11 (2012): 2108-2120.

[45] Fischer, Andreas; Frinken, Volkmar; Bunke, Horst; Suen, Ching Y. "Improving hmm-based keyword spotting with character language models." 12th International Conference on Document Analysis and Recognition. (ICDAR), 2013.

[46] Kumar, G.; Govindaraju, V., "A Bayesian Approach to Script Independent Multilingual Keyword Spotting," 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2014, vol., no., pp.357, 362, 1-4 Sept. 2014

[47] Ranjan, V.; Harit, G.; Jawahar, C.V., "Document Retrieval with Unlimited Vocabulary," IEEE Winter Conference on Applications of Computer Vision (WACV), 2015, pp.741-748, 5-9 Jan. 2015

[48] J. Almazan, A. Gordo, A. Fornes, and E. Valveny, “Word Spotting and Recognition with Embedded Attributes." IEEE Transactions on Pattern Analysis and Machine Intelligence. vol.36, no.12, pp.2552,2566, Dec. 1 2014