IJDAR
DOI 10.1007/s10032-008-0067-3

ORIGINAL PAPER

Matching word images for content-based retrieval from printed document images

Million Meshesha · C. V. Jawahar

Received: 23 July 2006 / Revised: 26 July 2008 / Accepted: 28 July 2008

Abstract As large quantity of document images is getting archived by the digital libraries, there is a need for an efficient search strategies to make them available as per users information need. In this paper, we propose an effective word image matching scheme that achieves high performance in the presence of script variability, printing variation, degradation and word-form variants. A novel partial matching algorithm is designed for morphological matching of word form variants in a language. We formulate feature extraction scheme that extracts local features by scanning vertical strips of the word image and combining them automatically based on their discriminatory potential. We present detailed performance analysis of the proposed approach on English, Amharic and Hindi documents.

Keywords Matching word images · Feature extraction · Partial matching · Degradation models

1 Introduction

With storage becoming cheaper and imaging devices becoming increasingly popular, efforts are on the way to digitize and archive large quantity of multimedia data (text, audio, image and video). In response, extensive research is being carried out to make the digital content accessible to users through indexing and retrieval of relevant documents from such collection of images [3,25], text [6,9], video and audio [4,11]. Most digital libraries aim at archiving books that are not available online. Success of text image retrieval systems mainly depends on the performance of optical character recognition (OCR), which convert scanned document images into texts. For indigenous scripts of India, Ethiopia, etc., there are no robust OCRs that can successfully recognize printed text images of varying quality, size, style and font. Hence, we need an alternate approach for effective access to the large collections of document images. A promising direction is to search for relevant documents using only image properties, without explicit recognition.

We reported a preliminary study conducted in this direction [2], with some prior results on searching word images using the word spotting approach. The system accepts textual query from users. The textual query is first converted to image by rendering. We render a given text word by setting font, style and point size. Features are then extracted from this image, and search is carried out for retrieval of relevant documents. Results of the search are a set of document images that are sorted in accordance to their relevance to the query word.

This paper presents an efficient word image matching scheme that is crucial for content-based document image retrieval from printed text collection. The major contributions are the following.

- Designing an innovative matching scheme for grouping together morphological variants of a word in a given language without an explicit textual representation. We demonstrate effective use of the matching algorithm for word images, which is flexible enough to compensate for morphological variants of a word. We present the matching algorithm in Sect. 4.
- Proposing a feature extraction scheme that is invariant to fonts, styles, sizes and degradations. We combine automatically a set of local features based on their discriminatory potential. The feature extraction scheme is effective

M. Meshesha · C. V. Jawahar (✉)
Center for Visual Information Technology,
International Institute of Information Technology,
Hyderabad 500 032, India
e-mail: jawahar@iiit.net

for representation of word images as discussed in Sect. 5.3. We argue that good performance can be obtained with the help of the combined local features.

We undertake extensive experiment to evaluate the effectiveness of the proposed approach. Results are demonstrated on English, Hindi and Amharic languages. Retrieval effectiveness measures such as recall, precision and F score are used to analyze the performance of the feature extraction and matching schemes.

2 Matching and retrieval from document collections

With the emergence of many digital libraries, document images are gaining popularity as an information source. Hence, access to the contents of the document image database is an important and a challenging problem in document image processing. A comprehensive review on the indexing and retrieval methods for document images is presented by Doermann [9]. This review focuses on the research issues when the documents are represented as text with the help of OCRs. In addition, a survey is made available on current trends of document image retrieval in digital libraries [21]. The main attempts are grouped into retrieval without recognition and recognition-based retrieval.

A number of works have been reported for retrieval from OCR documents [5,13,14,26]. Retrieval performance is shown to deteriorate significantly over scanned data with a character error rate greater than 5%. A preprocessor is mostly followed to improve errors in OCR texts and also to augment queries with terms which can be derived from original scanned text. An advancement is significantly made for retrieval from OCR English texts, but non-English text is far from solved [5]. A combination of more accurate OCR and more robust IR is still required. As a result, there is an increasing focus on retrieval without recognition, especially from poor quality document collections where OCR engines mostly fail. Word level image matching and retrieval has been attempted for printed [7,19,22,27,28,30] and handwritten [15,23,31] documents.

For printed documents, Chaudhury et al. [7] proposed a method to query documents at word level by exploiting the structural characteristics of the Indian scripts. They employed geometric feature graphs for representation, and suffix trees for indexing the printed text. Trenkle and Vogt [28] presented a preliminary experimental result on word-level image matching, where a query word is expanded to include its font variations. For each word image, features are extracted and a matching technique is employed for measuring the similarity between keyword and candidate words.

An attempt for document image retrieval based on word spotting is presented in [30]. The proposed technique addresses web document image retrieval problem by applying exact word matching procedure for indexing documents and using word-images as queries for searching. Further work on retrieval from printed document images is also reported by [18,27]. In [18], an efficient scheme for indexing document images is proposed based on a content sensitive hashing scheme. After document images are segmented into characters, features, such as the vertical and horizontal traverse densities are extracted and an n-gram-based document vector is constructed to measure similarity between documents using the dot product technique. Lu and Tan [19] proposed an inexact string matching technique to handle the problem of heavily touching characters and kerning during document image retrieval. After word images are represented by a primitive string, the proposed technique matches partial word images to estimate how a word image is relevant to the other and decide whether one word is a portion of the other.

The retrieval of historical printed documents is also attempted in [1,16,22]. Konidaris et al. [16] proposed a technique for word spotting in old Greek early Christian manuscripts. The aim is to search for keywords typed by the user in a large collection of digitized historical printed documents in which the retrieval result is optimized by user feedback. A method for retrieval of Ottoman documents based on word matching is reported by Ataer and Duygulu [1]. A hierarchical matching technique is designed to find similar instances of the word images. The matching method consecutively tests length similarity and similarity of quantized vertical projection profiles for the entire words, as well as for the ascender and descender part of the words. The work of Marinai et al. [22] presented word-level indexing of printed documents in the seventeenth to nineteenth centuries. The proposed approach indexes homogeneous document collections by automatically adapting to documents that vary in scripts and font styles. Unsupervised character clustering is done using self-organizing maps (SOM).

Matching and retrieval of word images from handwritten manuscripts using the word spotting idea was initially proposed by Manmatha and Croft [20]. The approach segmented pages into words and matched these words to retrieve relevant ones to a given query. An attempt is also made to select suitable features for matching handwritten word images [23]. Some of the considered features are single-valued features (such as profiles and gray-scale variance) for which one scalar value is calculated per column in the original image. Others are multivariate features (such as Gaussian smoothing and Gaussian derivatives) for which multiple values are calculated per image column. The work of Rath and Manmatha [24] discussed the problem of matching handwritten words in historical documents. By evaluating different matching techniques, they proposed dynamic time warping (DTW) for matching and retrieval of handwritten document images. Jain and Namboodiri [15] reported

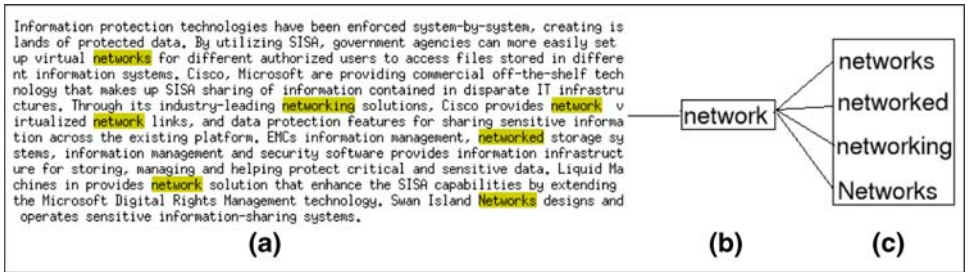

Fig. 1 **a** Sample real-life printed documents, **b** a particular root word, and **c** some of the word form variants in the document

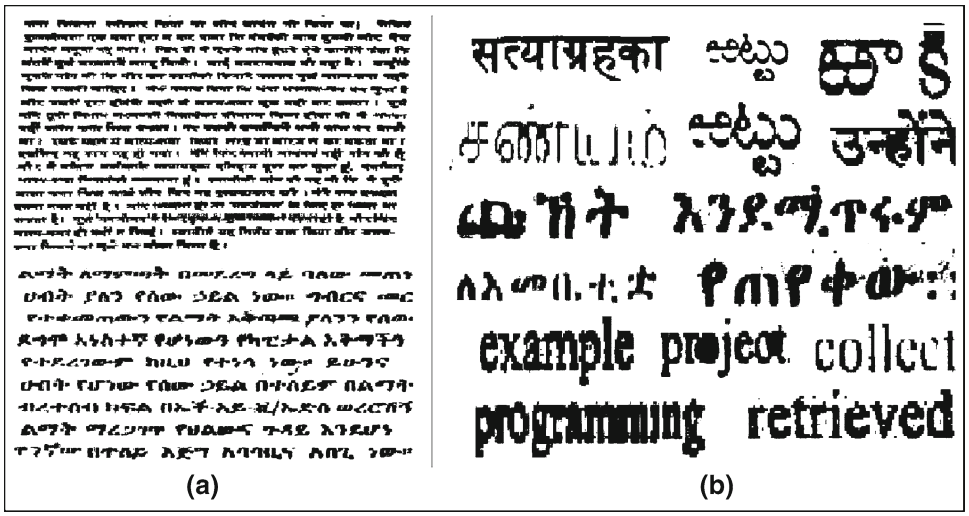

Fig. 2 **a** Samples of real-life printed documents, **b** some of the degraded words

DTW-based word-spotting algorithm for indexing and retrieval of online documents. Handwritten word image matching using gradient-based binary features is also presented [31]. By extracting binary features, correlation-based similarity measure is used for matching word images.

The main challenge with handwritten document retrieval is the writing variations. Correlation or DTW-based word image matching algorithms are designed to find exact matches to the query word and retrieve relevant documents in the presence of writing variations in the handwritten text. However, designing word-level matching scheme for searching in handwritten and printed document images faces many challenges and requires a thorough investigation and experimental analysis.

In this paper, we identify the various challenges in matching word images from printed documents as described below.

- *Language variability*: Each language has its own language rules, depending on which different morphological variants of a word are generated (see Fig. 1). To address this problem, we enable existing matching algorithm to perform morphological analysis in the image domain.
- *Degradation and printing variations*: Printed documents are often poor in quality. Figure 2 shows sample real-life document images, in which degradations like salt and pepper, cuts, blobs and erosion, could be observed. We need to simulate them so as to design and evaluate effective matching schemes for efficient retrieval from degraded document collections. Printed documents are also written in various fonts, styles and sizes. A good matching scheme will have to take-care of these variations.

In an attempt to develop a retrieval system in the image domain, much attention has been given to historical printed documents, where the OCR engine fails because of degradation and unseen fonts. Most approaches that are reviewed in this paper are useful for locating similar occurrences to the query word. Matching algorithms (such as DTW and correlation) are used for direct matching, which restricts searching for variants of a query word in the document collections. This way of searching has an adverse effect on the recall of the search engine.

The existence of language variabilities in a document results in the appearance of various word forms with the same meaning. As shown in Fig. 1, a word ‘network’ has many variants. When users query for document using this word they should retrieve documents that contain ‘networks’, ‘networked’, ‘networking’, etc. Users expect the document image retrieval system to provide the same convenience as it has been done effectively by text search engines. Most existing matching schemes fail to perform well in this situation. The ability to deal with these issues is crucial for effective matching scheme. We design partial matching scheme that is

flexible enough to control morphological variants of a word image.

3 Challenges for information retrieval from printed document images

Most printed books, manuscripts, newspapers, etc. that are archived in digital libraries are poor in quality, vary in printing and have word form variants. Indexing and retrieval of document images is a great challenge in this situation.

3.1 Modeling degradation in printed document

There are a number of artifacts in printed document images. Some of these are the following: (a) large ink-blobs joining disjoint characters or components, (b) cuts at arbitrary direction due to paper quality or foreign material, (c) erosion of black/white pixels from foreground/background, (d) degradation of printed text due to the poor quality of paper and ink, and (e) floating ink from facing pages.

Modeling document degradations enable us to closely examine the performance of matching scheme. We consider four categories of degradations that are commonly observed in printed documents. In this work, we model salt and pepper, cuts and blobs, in addition to erosion of boundary pixels simulated by Zheng and Kanungo [32]. Using these degradation models, we collect datasets of word images printed in Amharic, Hindi and English languages. Sample words are shown in Fig. 3. By varying parameter values, we generate datasets that mimic various levels of degradations in real-life, scanned document images. In Sect. 5, we use these datasets to analyze systematically the performance of features and matching schemes.

Printing variations: For printing in a specific script, there are different fonts, sizes and styles available for use. Hindi, Amharic and English use a number of fonts, each of them offering several stylistic variants and font sizes. These fonts, styles and sizes produce texts that greatly vary in their appearances (i.e. in size, shape, quality, etc.) in printed documents.

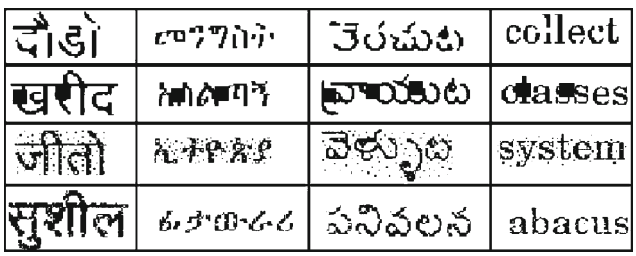

Fig. 3 Sample Hindi, Amharic, Telugu and English word images degraded with cuts, blobs, salt and pepper, and erosion of edge pixels (as shown from *top* to *bottom* row-wise)

We need to select suitable features that are invariant to printing variations.

3.2 Language variabilities

Different languages have their own language rules for formation of words from meaningful stems. On the basis of these rules, words with the same meaning may appear in various forms within texts. Hindi, Amharic and English are highly inflected languages which utilizes prefixes and suffixes to form words and to express grammatical relations. Words can be inflected for tense (present, future, and past), person (first, second, and third), gender (male and female), number (singular and plural) and/or case (direct, oblique, and vocative). Amharic and Hindi are verb-final languages, in which modifiers usually precede the nouns they modify. Unlike the word order, subject–verb–object in English sentences, the typical word order in Amharic and Hindi is Subject–object–verb.

English verbs are marked in the third person (he, she and it), like he sits. Most English verbs express tense through the use of various combinations of the auxiliary verbs ‘be’ and ‘have + main verb’. English has regular verbs that add ‘-ed’ or ‘-en’ to form the past tense, e.g., walk–walked, and irregular verbs that undergo internal vowel changes, e.g., drink–drank. English nouns are marked for numbers such as singular and plural. Possession is expressed by ‘-s’, e.g., mother’s.

Hindi verbs usually consist of a verb stem followed by an auxiliary verb, which is analogous to English word ‘going’, except that the order is reversed. Hindi verbs occur in the following forms: root (kha ‘eat’), imperfect stem (khat-A), perfect stem (khayA), and infinitive (khanA). The stems usually agree with nouns in gender and number. Tense distinction of present versus past is expressed by the auxiliary verb (honA ‘to be’). All nouns are assigned gender (masculine or feminine).

In Amharic, a verb form has one or more suffixes and prefixes, agreeing with the subject and object of the verb. Verb forms are derived by applying vowels and suffixes to the roots. Sometimes, consonants are geminated (doubled). There are at least ten different classes of verbs, each modifying its stem in a number of different ways. Amharic nouns take suffixes indicating possession, plural, and other grammatical functions. Amharic nouns are marked for genders, male and female. The female gender can be used to express smallness, e.g. ‘bet-itu’ (small house), bet-, ‘house’ and feminine marker -itu. Amharic nouns have singular and plural numbers. Plural is marked by the suffix ‘-oc’.

One of the main challenges in searching for relevant document using query words is the existence of word-form variants. The exact word that is present in the document being searched is a variant of the query word. As can be observed from language rules, there are two possible variants of a word: (i) word form variation and (ii) synonymous words

(words with the same meaning). For instance, for the query word 'direct', word-form variations include 'directs', 'directing', 'redirected', etc., where as, synonymous words include 'lead', 'organize' and 'guide' depending on the context. Addressing the second type of variation needs knowledge of the contextual meaning of the word and is beyond the scope of the present work. Here we address the first problem, namely, word-form variations.

In the text domain, most of the variations of word forms obey the language rules. A range of computational techniques are used for analyzing and representing written text in a language at one or more levels of linguistic analysis (such as morphological, syntactic, semantic, or pragmatic).

Morphological analysis identifies word-stem from a full word-form with the assumption that word variants have similar semantic interpretations. In the text-domain, a number of stemming algorithms have been developed to reduce a word to its stem or root form. Text-based search engines use this information to identify word form variants. However, for document-image retrieval, this information is not directly usable. There is a need to design a matching scheme that identifies word variants independent of language specific cases.

4 Morphological matching in the image domain

Language variabilities with in a document are unavoidable as they are constructed based on language rules. We attempt to mimic morphological analysis for matching a given word against its variants in the image domain.

4.1 Feature extraction methods

Feature extraction is the problem of gathering information from raw data, which is most relevant for a given application. Many different features have been employed in image processing and pattern recognition for representation of document images [10,12]. We experiment here with three categories of features: word profiles, moments and transform domain representations.

Word profiles provide a coarse way of representing word images for matching. Projection, transition, upper and lower profiles are features considered here for the representation of word images. While projection (F_1) and transition (F_2) profiles capture the distribution of ink along one of the two dimensions in a word image, upper (F_3) and lower (F_4) word profiles capture part of the outlining shape of a word. These features were employed for matching handwritten words [24]. To measure the distance between the upper and lower boundary of the word image, vertical distance (F_5) is used.

Moment-based features are computed for analyzing the shape of word images, where each order of moment has different information for the same image. In this paper, we employ statistical moments [such as mean (F_6), standard deviation (F_7) and skew (F_8)] and region-based moments [such as the zeroth-order M_{00} (F_9) and first-order M_{01} (F_{10}) moments].

Statistical moments measure the central tendency and distribution of ink pixels in word images. Region-based moments represent the entire shape region by describing the considered region using the pixels contained in that region. The knowledge of these moments can be used to calculate central moments, like cm_{02} (F_{11}) that computes squared distance from the y component of the mean.

A compact representation of a series of observations is possible in a transform domain using fewer coefficients. In this paper, we use discrete Fourier descriptors (F_{12}) for word shape representation. We compute N Fourier descriptors, $G(i), i = 0, 1, 2, \ldots, N-1$, following the vertical strips of the word image, where N is width of the word.

These features are either local or global features [29]. Local features represent a distinguishable small area of a region by extracting a sequence of features from sub-parts of the object. Whereas global features consider all points to describe some characteristics of the object such that one value is computed for representation.

4.2 Matching algorithms

Good matching performance can be achieved by a technique that aligns and finds the best match between pairs of queried and referenced word images [24]. In this respect, matching algorithms, such as cross correlation and dynamic time warping (DTW), have a great advantage in aligning and comparing word images with different sizes. Cross correlation is a statistical measure of how closely two signals are related [8].

The correlation function compares sequences of feature vectors of two word images by aligning them against each other, and at each position the correlation coefficient between the corresponding pairs of feature vectors of the two word images is computed. Optimal matching score gives a dissimilarity measure for the two aligned images. For decision purpose, the maximum of the resulting correlation score defines the best match.

Dynamic time warping (DTW) is a dynamic programming based approach [24] which is used to align sets of sequences of feature vectors and compute a similarity measure. Let two word images (say their moment) be represented as a sequence of features $\mathcal{G} = \mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_M}$ and $\mathcal{H} = \mathbf{H_1}, \mathbf{H_2}, \ldots, \mathbf{H_N}$. The DTW-cost between these two sequences is $D(M, N)$, where M and N are the lengths of the two sequences, which is calculated using the recurrence equation:

$$D(i, j) = \min \begin{cases} D(i-1, j-1) \\ D(i, j-1) \\ D(i-1.j) \end{cases} + d(i, j) \qquad (1)$$

where $d(i, j)$ is the cost in aligning the ith element of $\mathcal{G}$ with the jth element of $\mathcal{H}$. Squared Euclidean distance is used to compute $d(i, j)$. The use of $D(i, j-1)$, $D(i-1, j)$ and $D(i-1, j-1)$ in the calculation of $D(i, j)$ realizes a local continuity constraint. The optimal warping path (OWP) is the cheapest one with minimum distance, among all available paths in the DTW matching space, that runs from $D(0, 0)$ to $D(M, N)$ with length L. It is defined as

$$OWP(G, H) = \arg\min_{i} \ \text{cost}(W_i) \tag{2}$$

where W_i is a specific warping path of length L in the matching space. The matching cost $D(M, N)$ is normalized by the length of the optimal warping path to reduce the effect of word size variations during alignment.

The scenario of alignment of sequences of features of two words using DTW is shown in Figs. 4a and 5a. Once the matching process of two words is completed, we recover the optimal cost path from DTW matching space. A simple plot for matching profiles of two words is shown in Figs. 4b and 5b. The scenario depicts the optimal warping path (OWP) recovered with least matching cost.

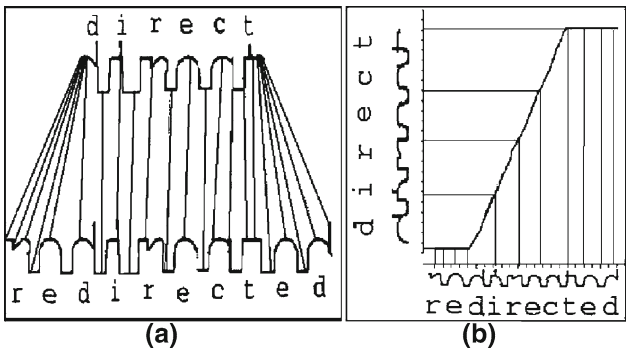

Fig. 4 Matching words using DTW. **a** Alignment of upper profiles of two English words, **b** the optimal matching path

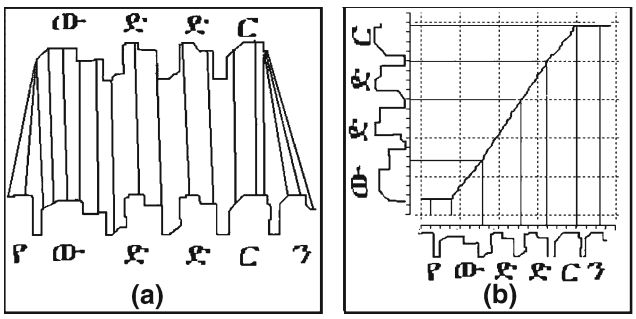

Fig. 5 Matching Amharic words using DTW. **a** Alignment of upper profiles of two words, **b** the optimal warping path

Algorithm 1 Morphological Matching

1: Partition the warping path ((with length L)) into three regions of length k ($L/3$) each: beginning, middle and end.
2: **for** $i = 1$ to k **do**
3: if there is matching cost concentration at the beginning.
4: reduce extra cost from the total matching score.
5: else break.
6: **end for**
7: **for** $i = L$ down to $2k$ **do**
8: if there is matching cost concentration in the end.
9: reduce extra cost from the total matching score.
10: else break.
11: **end for**
12: Normalize the matching score by the length of the optimal warping path.

4.3 Morphological matching to detect word forms

Most existing matching algorithms might be efficient for searching an exact match to a query word. However, the retrieval process for a good search engine is more involved than the simple word-level matching. Due to language variability, a word, with similar meanings, usually appears in various forms. This requires a method to conflate morphological variants of a word into a common stem [17]. In the text domain, variation of word forms may obey the language rules. Text search engines use this information to identify variants of a word and group them together. However, for document-image retrieval, this information is not directly usable.

We propose DTW-based partial matching technique that takes care of word form variations in the beginning and at the end. We follow Algorithm 1 for grouping word form variations. A close observation of the OWP shows that for word variants the DTW path deviates from the diagonal either in the horizontal or in the vertical directions from the beginning or end of the path. Note that, as the DTW path deviates from the diagonal line, the matching cost increases tremendously. For example, Fig. 4b presents the matching scenario of the root word ('direct') with its variant ('redirected'), in which the optimal path deviates from the diagonal line at the two extreme ends. This is because of the extra characters 're' and 'ed' in the word 'redirected'. Profiles of these extra characters have no or minimal contribution for measuring similarity of the two words and, hence their cost needs to be cut from the total matching score.

To this end, we analyze whether the dissimilarity between words is concentrated at the end, in the beginning or both. This involves partitioning the optimal warping path into three regions (beginning, middle and end) and checking the matching score at the two extremes (beginning and end). In Fig. 4b, it is observed that characters 're' in the beginning and 'ed' at the end of the word 'redirected' are get matched with characters 'd' and 't' of the word 'direct'. Due to this, there is

an alignment concentration at the beginning and end (see Fig. 4a), such that local matching cost is very high relative to other points. This increases the total matching score of the two words. We ignore the additional cost at the two extremes, reducing the total matching score. For instance, for a query word 'direct', the matching scores of the referenced words 'indirect' and 'redirected' are only the matching of the six characters, 'd-i-r-e-c-t', of both words.

Once an optimal sub-path is identified, a normalized cost corresponding to this segment is considered as the matching score for the pair of words. With the reduction of dissimilarity cost of a variant word (against its root word), we found that a large set of words get grouped together. This increases recall of the system during searching and retrieval.

The time complexity of the algorithm is $\theta(n)$, where n is the length of OWP of two words. This happens when variants of a given word are encountered, which occurs rarely in the collection. In most cases, no variants of a word exist and, hence the algorithm runs in constant time, which makes it practically applicable for searching in document images.

5 Results and discussion

In this section, we present an extensive experimental results to evaluate the performance of the feature extraction and matching schemes on clean and more on degraded datasets that varies in language and printing. The experiment is undertaken systematically; first, on clean and then on degraded data so as to evaluate relevance of the feature extraction scheme, and, then by mixing word images that are degraded and varies in printing so as to measure effectiveness of morphological matching scheme.

5.1 Datasets and performance measure

Performance analysis is greatly dependent on the dataset used for the experiment. We prepare ground-truth word-level datasets in English, Hindi and Amharic languages. From each language more than 4,000 basic word images and, an average of ten variants of each of these words are collected. These words are degraded using the four degradation models and are also duplicated using the specified fonts, sizes and styles in the language. With this we have a total of more than 800,000 word images for the three languages. We use these datasets to evaluate the performance of feature extraction and matching schemes proposed in the present work.

Performance is measured in terms of precision (P) and recall (R). While recall is the percentage of relevant words that are matched from the collection, precision is the percentage of matched words that are relevant. Figure 6 shows the effect of cutoff τ on recall and precision. The best possible cutoff point where recall and precision cross is set

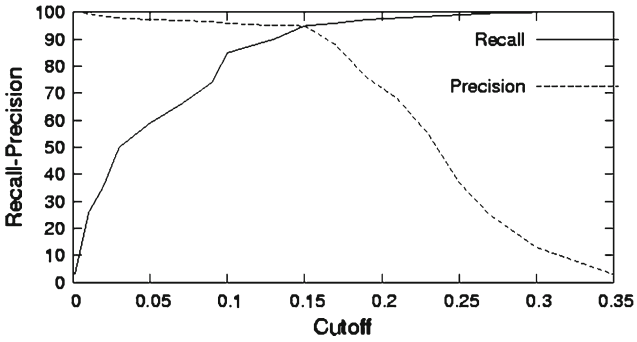

Fig. 6 Effect of cutoff point on recall and precision. Our interest is to arrive at a cutoff point τ where recall and precision cross

automatically by linear interpolation from recall and precision curves. Recall and precision are basically inversely related; an increase in recall is at the expense of precision and vice versa. To select features that balance recall and precision, we compute a single score using F measure (F) [17]. F measure (F) is a weighted harmonic mean of recall (R) and precision (P), which is expressed as $F = 2RP/(R+P)$.

Performance analysis shows that local features have better representational capability as compared to global features. If only global features are used to represent a word image, the local features of the word are mixed together. This results in mis-grouping of different word images, which share similar global features, while having different local features and contextual information. All the analysis given in the succeeding sections are limited to the local features.

Matching schemes: Using local features of word images, we evaluate both cross correlation and DTW matching algorithms. The average performance of the two algorithms is as follows. While DTW register 89.58% recall, 90.81% precision and 90.19% F score, cross correlation achieves 76.43% recall, 78.83% precision and 77.61% F score. The result reflects that DTW out-performs cross correlation in both recall, precision and F score. This is because cross correlation is relatively sensitive to small changes in the keyword being matched with referenced words though they are the same. DTW, on the other hand, compensates for most of the variations observed in printed word images during the alignment process. We therefore propose a modified DTW matching technique for computing similarity of word images in printed documents.

5.2 Feature selection for matching document images

Feature selection enables to explore the usefulness of selecting subsets of features that together have good predictive power. This requires a method to combine them and evaluate their performance on a given datasets. Given n features, a modified DTW algorithm is used to combine these features

Table 1 Average performance of local features for English, Hindi and Amharic languages

Features	Recall	Precision	F score
F_1	82.58	79.25	79.09
F_2	74.10	51.97	55.97
F_3	62.16	61.94	60.31
F_4	74.20	74.63	72.65
F_5	82.08	80.86	79.76
F_6	82.57	79.24	79.09
F_7	56.82	54.90	53.11
F_8	39.77	50.77	43.07
F_9	79.68	70.47	72.40
F_{10}	81.36	62.13	66.71
F_{11}	44.01	20.75	27.05
F_{12}	61.19	68.70	63.68

and compute matching score, as follows.

$$D(i,\, j) = \min \begin{cases} D(i-1,\, j-1) \\ D(i,\, j-1) \\ D(i-1.j) \end{cases} + \sum_{k=1}^{n} d^k(i,\, j) \qquad (3)$$

where $d^k(i,\, j)$ is the cost in aligning the ith and jth elements of feature k extracted from two words.

Table 1 shows average performance of individual features, in which vertical distance registers the highest F score of 79.76%. Projection profile and mean comes next with 79.09% each. Their relative good performance is attributed to the discriminatory information contained in them. Among the combined runs, highest F score is obtained by the following features. On English datasets, transition profile, lower profile, moment M_{00} and standard deviation register 90.05% F score. On Hindi, transition profile, lower profile, moment M_{00} and vertical distance result in 91.79% F score, and, on Amharic transition profile, lower profile, moment M_{00} and upper profile register 92.24% F score. For all languages combining transition profile, moment M_{00} and lower profile performs well. The difference lies on the fourth combined feature. For English standard deviation provides new data to control spatial distribution and spread of ascenders and descenders of the word. In Hindi words there is a headline *shirorekha*. By measuring the distance from this line to lower boundary, vertical distance supplies new information for better representation. Combining upper profile with lower profile is important to capture the outlining shape of the word both from the top and bottom so as to separate similar looking Amharic characters in a word. Thus, in general, more than 90% F score is obtained by combined features, which shows their high representational power than individual features. In the next section, we evaluate appropriateness of the combined features for representing clean and degraded word images that varies in printing and language.

5.3 Relevance of combined features for clean and degraded data

Quality, font, size and style difference are common in printed documents. We collect datasets of word images that are clean (but varies in font, size and style) and degraded using our degradation models. These words are represented using the combined features proposed in the previous section for experimentation.

Table 2 presents average performance of the proposed feature extraction scheme on clean data. The test result shows the effectiveness of our feature extraction scheme. On the average, 93.41, 96.63 and 85.73% F scores are registered for fonts, sizes and styles, respectively. The result for styles is reduced because of the complexities of words produced by italics typeface. By applying image processing better accuracy can be obtained for italics too. This shows that the combined local features suggested by the feature extraction scheme are invariant against printing variations.

The test result in Table 3 provides average performance of the combined features on degraded datasets. On the average, 92.72, 95.72, 89.53 and 93.76% F scores are obtained on

Table 2 Test results on clean datasets of the various fonts, sizes and styles of English, Amharic and Hindi languages

Item	English				Amharic				Hindi			
	Type	Rec	Pre	Fsc	Type	Rec	Pre	Fsc	Type	Rec	Pre	Fsc
Font	Ariel	96.74	97.62	97.48	PowerGeez	97.31	97.17	97.24	Yogesh	95.92	97.46	96.68
	Times	97.01	97.78	97.39	VisualGeez	92.22	93.04	92.63	Ganesh	81.09	85.27	83.13
	Courier	94.13	95.29	94.71	Alphas	93.86	96.93	95.37	Natraj	92.84	95.39	94.10
	SanSerif	83.15	91.83	87.28	Feedel	90.01	91.79	90.89	Surakh	93.89	94.21	94.05
Style	Normal	97.78	98.32	98.05	Normal	97.15	98.40	97.77	Normal	96.41	96.79	96.60
	Bold	96.53	95.07	95.79	Bold	96.28	97.33	96.80	Bold	94.24	94.78	94.51
	Italic	58.19	61.96	60.02	Italic	76.51	71.92	74.14	Italic	57.03	58.76	57.88
Size	10	96.45	96.80	96.63	10	96.47	97.28	96.87	10	93.29	96.26	94.75
	12	97.12	98.41	97.76	12	97.36	98.54	97.95	12	96.18	96.84	96.51
	14	97.29	97.16	97.22	14	95.37	97.67	96.51	14	94.42	96.58	95.49

Rec recall, *Pre* precision, *Fsc* F score

Table 3 Test results on degraded word images of the three languages: English, Hindi and Amharic

Degradation	English			Hindi			Amharic		
	Recall	Precision	F score	Recall	Precision	F score	Recall	Precision	F score
Cuts	94.12	88.73	91.35	92.34	92.41	92.37	93.58	95.29	94.43
Salt and pepper	96.64	96.98	96.81	93.28	93.17	93.20	96.87	97.41	97.14
Blobs	89.24	87.07	88.14	85.95	92.33	89.03	89.72	93.33	91.49
Erosion	92.61	93.85	93.23	92.77	92.58	92.67	95.06	95.69	95.37

Table 4 Performance of partial matching technique; test results are shown both before and after incorporating the partial matching module

Item	Type	Before			After		
		Recall	Precision	F score	Recall	Precision	F score
Font	Arial	90.77	93.28	92.01	97.52	99.67	98.58
	Courier	90.61	93.32	91.95	96.59	99.33	97.94
	SansSerif	63.09	87.93	73.47	91.46	96.77	94.04
	Times	91.46	90.87	91.16	97.08	99.21	98.13
Size	10	88.56	91.64	90.07	97.48	99.14	98.50
	12	91.10	93.17	92.12	97.52	99.67	98.58
	14	80.49	88.26	84.20	96.36	98.62	97.48
Style	Normal	90.68	93.22	91.93	97.52	99.67	98.58
	Bold	82.24	87.53	84.80	96.79	99.51	98.13
	Italic	56.72	55.85	56.28	76.20	81.98	78.98
Degradation	Cuts	83.57	87.02	85.26	91.83	96.26	93.99
	Blobs	76.14	82.18	79.05	88.71	96.04	92.23
	Salt and Pepper	92.91	92.74	92.83	93.28	99.19	96.14
	Erosion	90.46	90.38	90.42	94.46	97.34	95.88

cuts, salt and pepper, blobs and erosion, respectively. Minimum result is registered for blobs, because word images degraded by blobs are relatively confusing to identify as this noise adds more ink pixels that changes the visual meaning of a word image. The result registered language-wise indicates that, on the average, 92.38, 91.82 and 94.61% F scores are obtained for English, Hindi and Amharic, respectively. From the result we can observe the following. First, high performance is registered on degraded word images. Second, results are comparable across the three languages. This means that the proposed feature extraction scheme are insensitive to degradations, even with a change in script.

5.4 Effectiveness of morphological matching

As a solution to the problem of language variability we propose a partial matching scheme. We conduct experiment in two steps to evaluate its performance. First, the matching process is done skipping the partial matching module. We compute recall, precision and F score for this. Then, the test is repeated incorporating the partial matching procedure. Table 4 presents comparison of the test result before and after applying the partial matching scheme. The result shows that an average improvement of 11.85% F score is registered. This indicates that the proposed partial matching algorithm is able to greatly enhance the performance of the system. As can be observed from Table 4, the algorithm registers in the nineties for most of the cases, except for italic font style. This is because of the complexity of the word shape created by this style for representation.

Precision versus recall graph is shown in Fig. 7. The effectiveness of our scheme is observed from the graph as it increases both precision and recall by moving the entire curve up and out to the right. This shows a significant contribution of the morphological matching procedure in grouping together the various variants of a word which have a great impact on the overall performance of the retrieval system.

The proposed word image matching scheme works well on diverse datasets. It registers more than 90% F score in most cases. This guarantees the advantage of the morphological

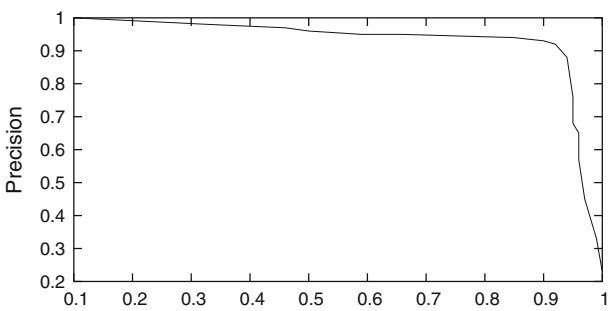

Fig. 7 Precision-recall graph for the proposed matching scheme

matcher to realize an effective search and retrieval system for printed document images that have many functionality. The success of this work greatly depends on the performance of the page segmentation algorithm and efficiency of the indexing scheme we are going to design. The matching scheme also needs to be more general so that it also addresses synonymous word variants that are seen in real-life documents.

6 Conclusions

Developing an efficient searching engine in the image domain requires designing an effective word image matching scheme. We achieved this by designing appropriate feature selection and morphological matching schemes in the image domain. By organizing datasets that encompass the various scripts, fonts, styles, sizes, degradations and word forms, we experiment the efficacy of the proposed approach. Performance analysis reveals that the use of combined local features is effective to manage the various artifacts and printing variations seen in printed documents. In addition, the introduction of DTW-based partial matching scheme enables to control morphological variants of a word for performance improvement. Having an effective partial matcher is crucial for developing a sophisticated data structure for indexing and clustering word images. This is one of the major steps for comprehensive performance analysis in real-life documents.

References

1. Ataer, E., Duygulu, P.: Retrieval of ottoman documents. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 155–162 (2006)
2. Balasubramanian, A., Meshesha, M., Jawahar, C.V.: Retrieval from document image collections. In: Proceedings of the Seventh International Association for Pattern Recognition (IAPR) Workshop on Document Analysis Systems (DAS), pp. 1–12 (2006)
3. Breiteneder, C., Eidenberger, H.: Content-based image retrieval in digital libraries. In: Proceedings of International Conference on Digital Libraries: Research and Practice, pp. 67–74 (2000)
4. Brown, M., Foote, J., Jones, G., Jones, K.S., Young, S.: Open-vocabulary speech indexing for voice and video mail retrieval. In: Proceedings of the Fourth ACM International Multimedia Conference, pp. 307–316 (1996)
5. Callan, J., Kantor, P., Grossman, D. (eds.): Information retrieval and OCR: from converting content to grasping meaning. In: Proceedings of the SIGIR 2002 Workshop, University of Tampere, Finland, 15 August 2002
6. Chan, J., Ziftci, C., Forsyth, D.: Searching off-line arabic documents. In: Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1455–1462 (2006)
7. Chaudhury, S., Geetika Sethi, A.V., Harit, G.: Devising interactive access techniques for indian language document images. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), pp. 885–889 (2003)
8. Devillard, N.: Infrared jitter imaging data reduction algorithms. Astron. Soc. Pac. Conf. Ser. **172**, 172–333 (1999)
9. Doermann, D.: The indexing and retrieval of document images: a survey. Comput. Vis. Image Underst. **70**(3), 287–298 (1998)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Willey, New York (2001)
11. Foote, J.: An overview of audio information retrieval. ACM Multimed. Syst. J. **7**, 2–10 (1999)
12. Gonzalez, W.: Digital Image Processing. Addison–Wesley, Massachusetts (1992)
13. Harman, D.K. (ed.): In: Proceedings of TREC-4. NIST Special Publication 500-236, Gaithersburg, MD, November 1995
14. Hawking, D.: Document retrieval in OCR-scanned text. In: Proceedings of the Sixth Parallel Computing Workshop (1996)
15. Jain, A.K., Namboodiri, A.M.: Indexing and retrieval of on-line handwritten documents. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), pp. 655–659 (2003)
16. Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., Perantonis, S.J.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. Int. J. Doc. Anal. Recognit. **9**(2), 167–177 (2007)
17. Korfhage, R.: Information Storage and Retrieval. Willey, New York (1997)
18. Kumar, A, Jawahar, C.V., Manmatha, R.: Efficient search in document image collections. In: Proceedings of 8th Asian Conference on Computer Vision (ACCV'07), Part I, LNCS, vol. 4843, pp. 586–595 (2007)
19. Lu, Y., Tan, C.L.: Information retrieval in document image databases. IEEE Trans. Knowl. Data Eng. **16**(11), 1398–1410 (2004)
20. Manmatha, R., Croft, W.B.: Word spotting: indexing handwritten archives. In: Maybury, M. (ed.) Intelligent Multimedia Information Retrieval Collection, pp. 43–64. AAAI/MIT Press, Cambridge (1997)
21. Marinai, S.: A survey of document image retrieval in digital libraries. In: 9th Colloque International Francophone Sur l'Ecrit et le Document (CIFED), pp. 193–198 (2006)
22. Marinai, S., Marino, E., Soda, G.: Font adaptative word indexing of modern printed documents. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **28**(8), 1187–1199 (2006)
23. Rath, T., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), pp. 218–222 (2003)
24. Rath, T., Manmatha, R.: Word image matching using dynamic time warping. Proc. Conf. Comput. Vis. Pattern Recognit. **2**, 521–527 (2003)
25. Rui, Y., Huang, T., Chang, S.: Image retrieval: Past, present, and future. J. Vis. Commun. Image Represent. **10**, 1–23 (1999)
26. Taghva, K., Borsack, J., Condit, A., Erva, S.: The effects of noisy data on text retrieval. J. Am. Soc. Inf. Sci. **45**(1), 50–58 (1994)
27. Tan, C.L., Huang, W., Yu, Z., Xu, Y.: Imaged document text retrieval without OCR. IEEE Trans. Pattern Anal. Mach. Intell. **24**(6), 838–844 (2002)
28. Trenkle, J.M., Vogt, R.C.: Word recognition for information retrieval in the image domain. In: Symposium on Document Analysis and Information Retrieval, pp. 105–122 (1993)
29. Trier, O.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition: a survey. Pattern Recognit. **29**(4), 641–662 (1996)
30. Zagoris, N.P.K., Chamzas, C.: Web document image retrieval system based on word spotting. In: IEEE International Conference on Image Processing, pp. 477–480 (2006)
31. Zhang, B., Srihari, S.N., Huang, C.: Word image retrieval using binary features. Proc. Doc. Recognit. Retr. **XI**, 45–53 (2004)
32. Zheng, Q., Kanungo, T.: Morphological degradation models and their use in document image restoration. In: International Conference on Image Processing, pp. 193–196 (2001)