Handwritten Text Retrieval from Unlabeled Collections

Santhoshini Gongidi$^{(\boxtimes)}$ and C. V. Jawahar

Centre for Visual Information Technology, IIIT Hyderabad, Hyderabad, India
santhoshini.gongidi@research.iiit.ac.in, jawahar@iiit.ac.in

Abstract. Handwritten documents from communities like cultural heritage, judiciary, and modern journals remain largely unexplored even today. To a great extent, this is due to the lack of retrieval tools for such unlabeled document collections. This work considers such collections and presents a simple, robust retrieval framework for easy information access. We achieve retrieval on unlabeled novel collections through invariant features learned for handwritten text. These feature representations enable zero-shot retrieval for novel queries on unlabeled collections. We improve the framework further by supporting search via text and exemplar queries. Four new collections written in English, Malayalam, and Bengali are used to evaluate our text retrieval framework. These collections comprise *2957* handwritten pages and over *300K* words. We report promising results on these collections, despite the zero-shot constraint and huge collection size. Our framework allows the addition of new collections without any need for specific finetuning or labeling. Finally, we also present a demonstration of the retrieval framework. [Project Page].

Keywords: Document retrieval · Keyword spotting · Zero-shot retrieval

1 Introduction

Digitized handwritten documents are a colossal source of information. Many communities use handwritten records like cultural heritage collections, judicial records, and modern journals to gather information. Many of these documents still remain to be made search-friendly. This is where handwritten search tools come in handy to make the textual content of these documents easily accessible. However, there is a lack of these search tools in digital libraries even today. For Latin scripts, document retrieval demonstrations are primarily experimental. Such demonstrations rely on transcribed pages from the document collection to make it retrievable. However, due to the vast number of collections, there is a need for annotation-free approaches to develop retrieval solutions.

Keyword spotting (KWS), a recognition-free approach, is becoming increasingly popular for document retrieval. KWS is defined as the task of identifying
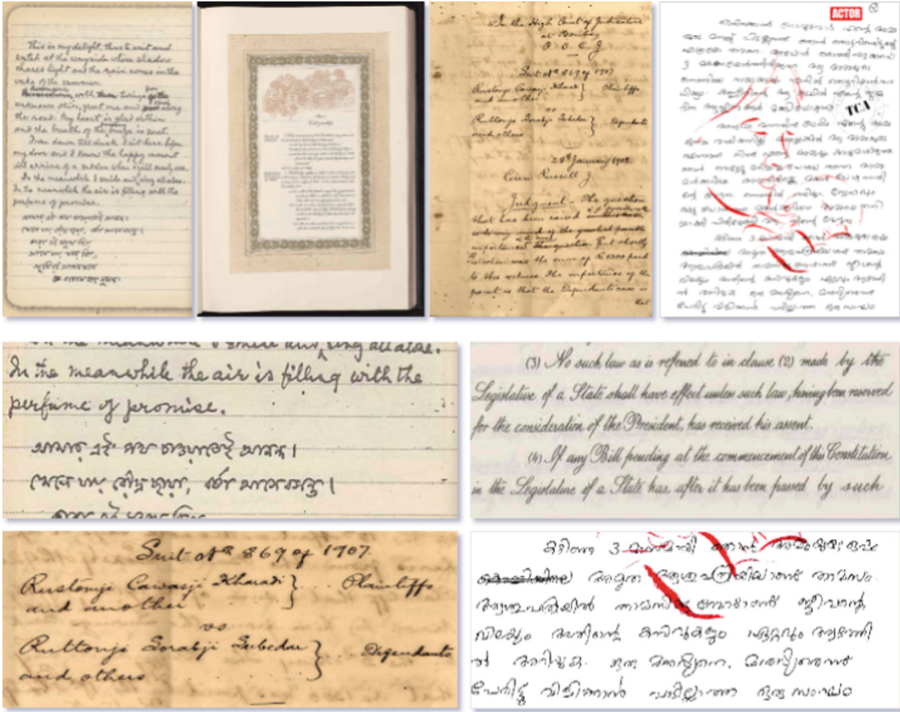
Fig. 1. Sample snippets from the historic and contemporary Indian handwritten collections: *Tagore's papers*, *Constitution of India manuscript*, *Bombay High Court judgements*, *Mohanlal writings*. Malayalam and Bengali scripts can be found in the bottom right and top left images, respectively. Challenges in these collections involve poor image quality, ink bleed, complex backgrounds, and layouts.

the occurrences of a given query in a set of documents. Existing state-of-the-art KWS solutions [2,10,14] are discussed majorly in the context of benchmark datasets like GW [12], IAM [11]. These datasets are neat and legible. It implies that the proposed approaches are evaluated on known handwriting styles and limited vocabulary sets. However, in practical settings, document retrieval tools must accommodate for new handwriting styles and open vocabulary. Additionally, the natural handwritten collections are associated with challenges like ink-bleed, illegible writing, various textures, poor image resolution, and paper degradation. This results in a huge domain gap between these collections and the benchmark datasets. This gap causes a drop in retrieval performance for unseen documents. Hence there arises a need to study the application of existing approaches for retrieval in unexplored documents. With this work, we propose an embedding-based framework that enables search over unexplored collections. Our framework performs retrieval on unseen collections without any need for finetuning or transcribed data. In other words, we perform zero-shot retrieval

from unexplored collections. This is achieved by learning holistic embeddings for handwritten text and text strings. These embeddings are also invariant to writing styles, paper textures, degradation, and layouts.

Another limitation of existing works is the lack of studies on document retrieval from non-Latin collections. We study and evaluate the performance of our framework on untranscribed collections in English and 2 Indic scripts: Malayalam and Bengali. This setup can easily be extended to other Indic scripts. Handwritten document collections from public digital libraries are used to demonstrate the efficiency of our retrieval pipeline. These collections contain a total of 2,957 pages and 313K words. Figure 1 shows snippets from chosen collections discussed in this work. Search on these specific records is difficult as they come with significant issues like poor image resolution, unstructured layouts, different scripts on a single page, ink-bleed, ruled and watermark backgrounds. We address these challenges and discuss a retrieval approach developed for these unseen collections. We utilize complementary modules along with spotting approaches to tackle some of these challenges. Another crucial aspect of any retrieval framework is its scalability. Therefore, we study and discuss efficient approaches to improve retrieval speed and memory requirements. We perform experiments on these collections to study the efficiency of our retrieval pipeline for searching across large document collections.

To the best of our knowledge, this work is an initial effort to perform handwritten search in a zero-shot setting at a large scale in multiple scripts. Our framework also supports querying with both textual and exemplar queries. We present a real-time handwritten document retrieval demonstration[1] for handwritten collections from India. The major contributions of this work are as follows:

i. An embedding based framework for retrieval from unseen handwritten collections.
ii. Study and discussion on the approaches towards generic keyword spotting that can handle unseen words and unseen writing styles.
iii. Evaluation of efficient indexing, and query matching methods.
iv. A demonstration to validate our end-to-end retrieval pipeline on 4 collections written in English, Bengali and Malayalam.

2   Related Works

In this section, we discuss earlier works in KWS and existing end-to-end retrieval demonstrations for handwritten collections.

Keyword Spotting: Earlier works in KWS use various methods such as template matching, traditional feature extraction along with sequence matching, and feature learning methods. Works [1,5] survey numerous spotting techniques and discuss commonly adopted methods for document retrieval. Recently popular

---

[1] Demo links available at project page.

feature learning methods for KWS aim to learn a holistic fixed-length representation for word images. Current top trends that use this approach are: attribute learning in PHOCNet [14], verbatim and semantic feature learning in triplet CNN [17], and joint representation learning in HWNet [10]. All these approaches are highly data-driven and can be categorized as segmentation-based approaches. Segmented word images from collections are a prerequisite in the above methods to train and extract features. In segmentation-free techniques, the network identifies potential text regions and then spots a given keyword. Wilkinson et al. [18] propose one such method for neural word search. An end-to-end network comprising a region proposal network and a holistic feature learning network is presented in their work. In another segmentation-free approach proposed for Bengali script [4], a CNN trained to identify the class label of word images is used to spot keywords in a dataset of 50 pages. Due to limited class labels and the incapability to handle out-of-vocabulary words, this setup is not suitable for spotting in new and unseen collections. In this work, we adopt a segmentation-based feature learning approach for searching. This choice enables us to ensure easier indexing and reasonable accuracy on new collections.

Retrieval Demonstrations: Search engine for handwritten collections was first introduced in [12] for George Washington's letters containing 987 document images. Recently, another concept referred to as probabilistic indexing(PrIx) has been introduced. Multiple large-scale Latin collections are indexed using PrIx, such as Bentham papers [15], Carabela manuscripts [16], and Spanish TSO collection [15]. A case study presented in [18] discusses a segmentation-free EBR for Swedish court records consisting of 55K images. These existing demonstrations are developed explicitly for Latin scripts. No such retrieval demonstrations are available for non-Latin scripts. Existing methods rely on transcribed pages from the collection to make it retrievable. The number of pages to be transcribed varies for different methods. 558 and 1213 transcribed pages are used to build demonstrators for the Carabela manuscripts and Bentham papers. In contrast, the demonstrator for Swedish court records used only 11 transcribed pages. The transcription costs are unfeasible as there are millions of massive collections written in different scripts. Our proposed pipeline performs retrieval on unseen collections without any need for finetuning. At the same time, our collections are equally complex when compared to the above mentioned collections. As the pipeline is not fine-tuned for specific writing styles or vocabulary of a new collection, our work is closely related to zero-shot learning.

3 Proposed Framework

In this section, we discuss the setup employed for searching in a document collection. Figure 2 shows an overview of the retrieval pipeline. Given a historic collection, the document images are forwarded to a text detector. The output from the detector is processed to extract text regions from the images. Valid text regions are selected and forwarded to the embedding network to compute

holistic embeddings. The computed embeddings and positions are used to create embedding index and position index for a collection. Pretrained networks are employed for embedding network and text detector. Further details about the modules are mentioned below.
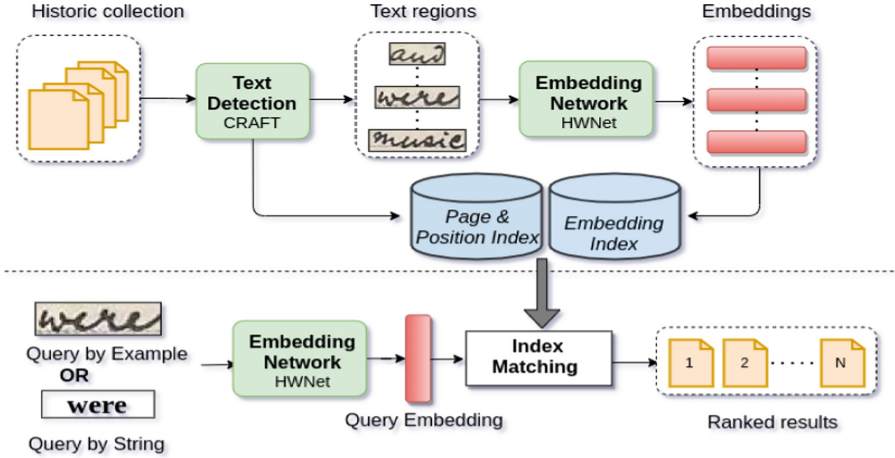

Fig. 2. Overview of the spotting and retrieval framework. (Top) Processing and indexing a handwritten collection to enable search operations. (Bottom) The flow of a query across the retrieval framework to retrieve relevant results.

The real-time flow of an input query is shown in the bottom image of Fig. 2. Text strings and exemplar images are supported as input query formats. These modes are referred to as QbS (Query by String) and QbE (Query by Example), respectively. Similarity search is performed using euclidean distance metric to retrieve relevant documents. Using brute force search in large indexes is costly due to a huge number of computations. We use approximate nearest neighbor approaches as these methods are non-exhaustive. This reduces the retrieval time for a given query and makes our framework efficient.

Characteristics: Our retrieval pipeline is designed so that the user can search within a specific collection with a single query. We believe that this is a reasonable assumption as the domains of the collection vary significantly and a user looking for specific information knows which collection to choose for querying. This choice enables to reduce the search index size and improve the retrieval time. Querying can be done through both text strings (QbS) and word image examples (QbE). The designed document retrieval pipeline is generic and can be extended to any collection. Our proposed setup is capable of dealing with unseen document images with reasonable confidence. Efficient representation learning enables this zero-shot retrieval from unexplored collections. In our setup, labeled handwritten data is only used during training the embedding network.

Embedding Network: Spotting a given keyword in an unseen document can be achieved by learning holistic feature representations for handwritten text. These representations must be invariant to writing styles, document degradation and poor image resolution. At the same time, the feature learning method should also generate discriminative representations for an open set of vocabulary. To achieve these goals, we employ the end-to-end deep embedding network, HWNetV2, discussed in [8]. This network learns a common subspace of embeddings for both word images and text strings. We utilize these embeddings as features for indexing and matching.
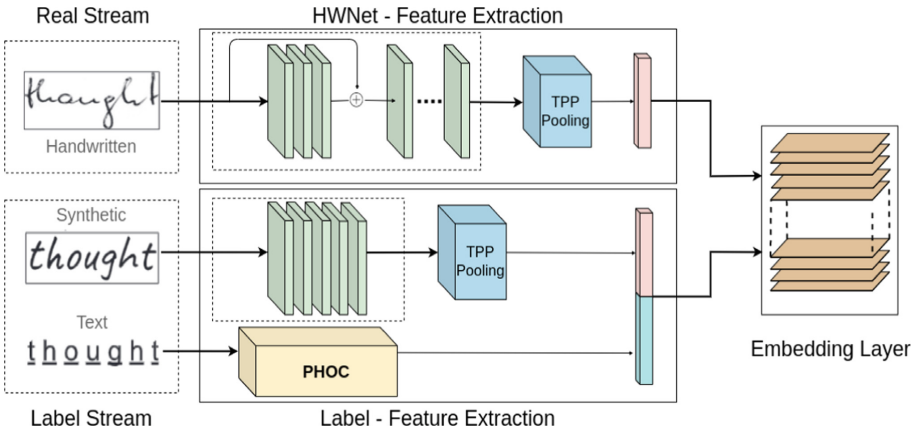

Fig. 3. Overview of the HWNetv2 embedding network for generating holistic handwritten text features. Embedding layer in the network is responsible for mapping the image features and text string features to a common subspace.

The overview of this architecture is shown in Fig. 3. The real stream, as shown in the figure, computes embeddings for word images from document collections using a deep network. Label stream is introduced to compute feature representations for text strings. It comprises of a PHOC [2] module and a shallow network. We request the readers to refer to [10] for technical details on architecture. The real stream and label stream embeddings are forwarded to an embedding layer. This layer learns to map these feature representations from two different modalities to a common embedding/representation space.

Feature Extraction: A collection is processed using Otsu thresholding to binarize and reduce the impact of different backgrounds and textures on the retrieval pipeline. We utilize a pretrained deep network CRAFT [3] to extract words from handwritten collections for text detection. This network is trained in a weakly supervised manner with pseudo ground truths on scene text datasets. Although CRAFT is trained on the scene text images, we observe reasonable detection rate for handwritten documents in Latin and Indic scripts. We believe that

the underlying idea behind the CRAFT detector makes the network robust to detect handwritten text. The detector works by localizing individual character regions and linking closer character regions to form words. As the detection network is used in a zero-shot setting, the text bounding boxes have a certain degree of error due to over-segmentation or under-segmentation. To overcome this, extreme affine augmentation is applied while training the HWNetv2 network. This is done to imitate segmentation issues and enable embedding learning for wrongly segmented inputs. This strategy helps to overcome segmentation issues during retrieval. The extracted text image regions are forwarded to the pretrained HWNetv2 to extract features.

Retrieval: The computed embeddings for a collection are indexed using inverted file index (IVF). While indexing a collection, the embeddings are clustered to identify Voronoi cells and their centroids, which are the representatives of the cells. For an input query, the computed embedding is used to search for top N similar Voronoi cells by matching the query embedding to the centroids. The identified Voronoi cells are further matched to all the embeddings in these top N cells to obtain top k closest embeddings. Documents corresponding to these embeddings are the top k relevant matches for the query. Only valid embeddings are indexed to reduce the index size. Invalid embedding corresponding to over-segmented, under-segmented text regions are avoided. Extremely small text regions are also pruned off as they correspond to stop words mostly. We use FAISS [7] to make efficient indexes and perform search operations. In our experiments, we cluster the embeddings into varying cell sizes depending on the total embeddings for a collection.

Query Processing: Input query is forwarded to the pretrained embedding network. Image queries are resized, normalized and forwarded to the HWNetv2 real stream. For textual queries, synthetic image is rendered containing the query string. This rendered image and the text string are forwarded to the HWNetv2 label stream. The computed query embedding is matched with relevant indexes. The index keys obtained from similarity matching are used to retrieve document ids, keyword position details from the page and position index. Relevant lines associated with the spotted query word are cropped from the corresponding documents using these positions. For this, the bounding box of the spotted query is extended along the image width and this selected area is presented as relevant lines for the input query.

4 Experiments and Results

4.1 Training Phase

The retrieval framework is developed for documents written in English, Bengali and Malayalam. We train the HWNetv2 network with existing handwritten datasets. IAM [11], GW [13] datasets are used for training English embeddings. For Bengali and Malayalam scripts, we use IIIT-INDIC-HW-WORDS [6]

dataset. Affine, elastic and color transformations are used to augment the training datasets. We also use the IIIT-HWS [9] synthetic dataset containing 1 million word images rendered using handwritten style fonts for pretraining the embedding network. The mean Average Precision (mAP) obtained on three training datasets are reported in Table 1. Note that the high evaluation scores are reported on rather clean, carefully collected datasets. This is unlike a real, practical setting where the handwriting, vocabulary and unseen layouts are prevalent.

Table 1. Evaluation metrics on the test split of IAM, IIIT-INDIC-HW-WORDS dataset. Evaluation done for both QbS and QbE settings. Full and OOV test refers to the complete test split and out-of-vocabulary test split respectively.

Dataset	QbS mAP		QbE mAP	
	Full test	OOV test	Full test	OOV test
IAM (English)	0.93	0.92	0.93	0.93
IIIT-INDIC-HW-WORDS (Malayalam)	0.99	0.98	0.99	0.99
IIIT-INDIC-HW-WORDS (Bengali)	0.95	0.95	0.95	0.94

4.2   Evaluation Datasets

Four collections from public websites and digital libraries are chosen to evaluate our pipeline: a collection consisting of handwritten poems and plays written by Rabindranath Tagore, a historic collection of Bombay High Court law judgements from the year 1864, a modern collection of handwritten blogs by an actor and a handwritten manuscript of Constitution of India (CoI). Source links for these documents are linked in Table 2. The table also lists the scripts used, total pages and estimated words in these collections. Figure 1 shows sample blocks from these collections. The pages in these collections have printed text, handwritten text, watermarks and backgrounds along with illegible text due to degradation, difficult writing styles. Our retrieval framework is capable of retrieving meaningful results for most of the queries despite these challenges. The low image resolution in CoI manuscript and Tagore's papers is also handled implicitly without the need for additional processing.

Table 2. Digital handwritten collections demonstrated and evaluated in this work. Digital sources are linked in the collection title column.

Collection title	Script	Pages	Words (est.)
Bombay High Court (BHC) records	English	1,964	207,028
Constitution of India (CoI) manuscript	English	479	42,040
Tagore's Papers (TP)	English, Bengali	125	10,990
Mohanlal Writings (LAL)	Malayalam	389	53,111

4.3   Evaluation and Discussion

In this section, we discuss the results obtained on the four collections mentioned above. Unlike the training datasets mentioned in Table 1, these collections are unexplored and unlabeled. Therefore, reporting exhaustive evaluation metrics is not feasible. Labeled data is a prerequisite to compute mAP. In place of mAP, we report the precision (P) of top-k ranked results at $k = 1, 10, 25$. This evaluation method is also followed in similar case study discussed in [18]. This web-scale metric does not require the knowledge of all relevant instances for a given query. We pick random queries from each of the collections and report the top-k precision (P_k) in Tables 3, 4, 5 and 6. The results are reported for QbS retrieval setting in these tables. We observe similar results for QbE query retrieval as well. The retrieval framework achieves similar performance for both seen and unseen out-of-vocabulary (OOV) queries. For queries like *vacancy* and *sudden* low values of P_k at $k = 10, 25$ are reported as these are rare words that are used in the collection. Evaluating on CoI manuscript for extremely rare OOV words like *Madras, Travancore,* and *surcharge,* the P_k for $k = 1$ is 1.00. The framework is also capable of retrieving related terms for a given query. For example, for the query *vacancy* relevant results include the term *vacant* as well. More such pairs are *discharge-surcharge, clause-subclause, constitution-constitute,* and *Vice President-President.*

Table 3. Precision for queries in *Constitution of India* collection. Unseen vocabulary during training time are marked in blue color.

	Constitution	India	Right	Vacancy	Rajpramukh	Deputy	Discharge
$P_{k=1}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$P_{k=10}$	1.00	1.00	1.00	0.60	0.80	1.00	1.00
$P_{k=25}$	1.00	0.96	0.92	0.24	0.76	1.00	0.88

Table 4. Precision for queries in *Bombay High Court records* collection. Unseen queries during training time are marked in blue color.

	Bombay	Court	August	Judge	Defendant	Registrar	Plaintiff
$P_{k=1}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$P_{k=10}$	0.90	1.00	1.00	1.00	1.00	0.90	1.00
$P_{k=25}$	0.88	1.00	0.88	1.00	1.00	0.92	1.00

Table 5. Precision for queries in *Tagore's papers*. Unseen queries during training time are marked in blue color.

	heart	light	তরে	sudden	thy	ছিল	অনুবাদ
$P_{k=1}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$P_{k=10}$	1.00	1.00	0.70	0.60	0.90	0.40	1.00
$P_{k=25}$	1.00	0.92	0.52	0.24	0.88	0.16	0.64

Table 6. Precision for queries in *Mohanlal writings*. Unseen queries during training time are marked in blue color.

	സിനിമ	ജീവിതം	കുടുംബ	മോഹൻലാൽ	നടൻ	ഇന്ത്യ
$P_{k=1}$	1.00	1.00	1.00	1.00	1.00	1.00
$P_{k=10}$	1.00	1.00	0.50	1.00	0.70	1.00
$P_{k=25}$	1.00	1.00	0.20	0.44	0.28	0.72

We also show the qualitative results obtained for the four collections in different settings. Figures 4, 5 and 6 shows the top retrieval results for both seen and unseen queries. Our framework retrieves accurate results despite the drastic style variations. For example, positive results shown for 3 English handwriting styles in both Fig. 4 and Fig. 5. For the query *registrar* shown in Fig. 4(a), the retrieval results contain both printed text images and handwritten images. It shows the robustness of the learnt embeddings. Without any prior training for these specific handwriting styles, the retrieval results are promising.

Fig. 4. Top-10 qualitative search results from collections: BHC records and CoI manuscript. Queries are shown at the top and incorrect retrievals are highlighted in red. Querying is done in QbS setting. Poor image quality of the collections is also shown here. (Color figure online)

(a) Top-5 matches   (b) Rare OOV query results

Fig. 5. Qualitative search results from Tagore's papers collection for QbS setting. (a) Top-5 results for English and Bengali queries. OOV queries are highlighted in blue and incorrect retrievals are highlighted in red. (b) Showing top-1 result for a few rare OOV words from the collection. (Color figure online)

(a) QbS Retrieval   (b) QbE Retrieval

Fig. 6. Qualitative search results from Mohanlal's writings collection for QbS and QbE setting. Top row shows the queries. OOV queries are highlighted in blue and incorrect retrievals are highlighted in red. (Color figure online)

In Figs. 5 and 6, the incorrect retrievals match closely to the query keyword. We also show promising results in QbE retrieval mode for the Malayalam collection in Fig. 6(b). Despite the watermarks present in the images, the obtained results are accurate and relevant. QbE mode of querying is especially helpful while searching for signatures and unknown symbols encountered in a collection.

4.4   Comparative Results for Retrieval

This section discusses the time complexity and memory requirements of the indexing and retrieval module. We compare our index matching algorithm to two other approaches using K-D Tree and nearest neighbour approach. Nearest neighbour approach involves an exhaustive search over all the samples in a

given index. K-D Tree and IVF based search are both non-exhaustive. Search operations are conducted in selective lists to obtain retrieval results. We compare and report the retrieval time and the mAP obtained for these three index matching approaches in Table 7. We also report the index sizes(disk memory) for these methods. The retrieval time and index size are obtained by indexing and querying the fairly large BHC records collection. As mAP evaluation metric requires ground truth, we utilize the IAM test set to report this metric for the search methods. The search method used in this work is both efficient in terms of both retrieval time and QbS mAP. The retrieval algorithm used in this work decreases the time by 86% compared to the simple nearest neighbour approach. The choice of retrieval algorithm is really important when the indexes are huge. Therefore, we also study the effect of index sizes on the search time. We observe that time required to perform brute-force search increases rapidly with an increase in index size compared to the inverted index based approach. Therefore, we use the inverted index based approach for searching.

Table 7. Retrieval time vs. Accuracy for query matching algorithms. Search operations performed in an index with 200K samples with feature dimension as 2048.

Search method	Exhaustive	Retrieval time (sec/query)	Index Size (GB)	QbS mAP (IAM test)
K-D tree	No	0.59	3.5	0.95
Nearest neighbour*	Yes	0.15	1.6	0.96
This work*	No	**0.02**	1.6	0.95

*Note: * Computed using FAISS library.*

5   Conclusion

Document retrieval from complex handwritten records containing an unseen vocabulary set is challenging. In this work, we discuss a document retrieval framework for unseen and unexplored collections. This framework can perform search operations on these collections without any fine-tuning for a specific collection. We discuss and evaluate our method both quantitatively and qualitatively on four different collections written in three scripts, of which two are Indic. Our framework performs reasonably well on these sizeable collections. Even for rare words, the top results are accurate. Utilising this simplified framework we plan to introduce more Indic collections written in other Indic scripts. Finally, we also present a demonstration to showcase the usefulness of the retrieval framework.

References

1. Ahmed, R., Al-Khatib, W.G., Mahmoud, S.: A Survey on handwritten documents word spotting. Int. J. Multimed. Inf. Retr. **6**(1), 31–47 (2016). https://doi.org/10.1007/s13735-016-0110-y
2. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. TPAMI **36**, 2552–2566 (2014)
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: CVPR (2019)
4. Das, S., Mandal, S.: Keyword spotting in historical Bangla handwritten document image using CNN. In: ICACCP (2019)
5. Giotis, A.P., Sfikas, G., Gatos, B., Nikou, C.: A survey of document image word spotting techniques. PR **68**, 310–332 (2017)
6. Gongidi, S., Jawahar, C.V.: IIIT-INDIC-HW-WORDS: a dataset for Indic handwritten text recognition. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12824, pp. 444–459. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86337-1_30
7. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
8. Krishnan, P., Dutta, K., Jawahar, C.: Word spotting and recognition using deep embedding. In: DAS (2018)
9. Krishnan, P., Jawahar, C.: Generating synthetic data for text recognition. arXiv preprint arXiv:1608.04224 (2016)
10. Krishnan, P., Jawahar, C.V.: HWNet v2: an efficient word image representation for handwritten documents. IJDAR **22**(4), 387–405 (2019). https://doi.org/10.1007/s10032-019-00336-x
11. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. IJDAR **5**, 39–46 (2002). https://doi.org/10.1007/s100320200071
12. Rath, T.M., Manmatha, R., Lavrenko, V.: A search engine for historical manuscript images. In: SIGIR (2004)
13. Rath, T.M., Manmatha, R.: Word spotting for historical documents. IJDAR **9**, 139–152 (2007). https://doi.org/10.1007/s10032-006-0027-8
14. Sudholt, S., Fink, G.A.: PHOCNet: a deep convolutional neural network for word spotting in handwritten documents. In: ICFHR (2016)
15. Toselli, A.H., Romero, V., Sánchez, J.A., Vidal, E.: Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing. In: ICDAR (2019)
16. Vidal, E., et al.: The CARABELA project and manuscript collection: large-scale probabilistic indexing and content-based classification. In: ICFHR (2020)
17. Wilkinson, T., Brun, A.: Semantic and verbatim word spotting using deep neural networks. In: ICFHR (2016)
18. Wilkinson, T., Lindström, J., Brun, A.: Neural word search in historical manuscript collections. ArXiv (2018)