Content Level Access to Digital Library of India Pages

Praveen Krishnan* Ravi Shekhar* C.V. Jawahar

Center for Visual Information Technology, IIIT Hyderabad, India
{praveen.krishnan, ravi.shekhar}@research.iiit.ac.in, jawahar@iiit.ac.in

ABSTRACT

In this paper, we propose a framework for content level access to the scanned pages of Digital Library of India (DLI). The current Optical Character Recognition (OCR) systems are not robust and reliable enough for generating accurate text from DLI pages. We propose a search scheme which fuses noisy OCR output and holistic visual features for content level access to the DLI pages. Visual content is captured using Bag of Visual Words (BoVW) approach. We show that our fusion scheme improves over the individual methods in terms of mean Average Precision (mAP) and mean precision at 10 (mPrec@10). We exploit the fact that OCR has a high precision while BoVW has a high recall. We use a modified edit distance to improve the order of results ranked by BoVW. Experiments are carried out on large datasets of DLI pages in Hindi and Telugu languages. We validate our method on more than 10,000 pages and 4 Million words, and report a mAP of around 0.8 and mPrec@10 of more than 0.9. We show improvements over BoVW by introducing query expansion. We also demonstrate a textual query interface for the search system.

Keywords

Content Level Access, BoVW, OCR, DLI

1. INTRODUCTION & RELATED WORK

Digital Library of India (DLI) has emerged as one of the largest collections of document images in Indian scripts [1]. DLI, as a part of Million Book Project (MBP), has contributed to the free access of knowledge to Billions of people. In addition, it also helped in digitally archiving the rare and precious books in many of the Indian languages. All these digital contents are stored as scanned images of printed documents. A major challenge presently faced by the DLI is the lack of content level access to the individual pages. As it stands, content (in Indian languages) can be primarily accessed only by the meta data of the books, which is manually created.

Figure 1: Examples of retrieved words in Hindi for the query 'SWATANTR' from DLI books considered for the experimentation. Note that many of these words are degraded and challenging for an OCR

DLI archives printed content (mostly books and journals) as scanned document images. Since robust OCRs are not readily available for reliably converting them into text, they are not directly accessible for a search engine. Objective of this work is to demonstrate the feasibility of content level access to a large collection of scanned books from DLI. These books are available as binary images, and the binarization has already resulted in significant cuts and merges in the word images. We argue that, even if a reliable high performance OCR is not readily available, effective search is still possible. Figure 1 shows an example set of retrieved words for a query 'SWATANTR'. It can be observed that many of these words are highly degraded and challenging, even for a commercial English OCR (if the script is English).

Since OCRs are not an immediate feasibility (see Section 3 for quantitative performance of OCR on DLI pages) for building search engines, we naturally move towards the image based search and retrieval techniques. Such methods are often formulated in a "query by example" setting. In such a setting, the query is a word image, and the search happens by finding similar word images by comparing the feature representations. Often this process is speeded up with an index structure. Though query by example is still an acceptable scheme for many natural image retrieval applications, typical users of a digital library prefer a textual interface for document image collections. One would like to search just as in "Google".

There has been significant progress in the area of image search in recent years. The text-annotated Internet image

*Equal contribution

ICVGIP '12, December 16-19, 2012, Mumbai, India

search engines have started to use the image features for various tasks. There have also been many successful attempts [14, 18] to retrieve specific objects from a large collection of images and videos. Many products such as Google goggles, SnapTel etc. have already reached the end users. We are inspired by the success of these methods.

In this work, our objective is to design techniques that perform reliably on degraded DLI images, and seamlessly scale up to large datasets. We demonstrate our methods on more than 100 books from two Indian languages - Hindi and Telugu. In the entire process, we also design techniques which exploit the noisy outputs of the present day OCRs. We also show that our implementation can retrieve relevant word images in a fraction of second from Millions of word images. In order to quantitatively validate the performance of the method, we manually annotate 1000 pages each in Hindi and Telugu. On these data sets, we measure the performance using mAP and on the rest un-annotated datasets (more than 100 books), we measure the mPrec@N by manual counting. We discuss our results in detail in Section 3.2. Before discussing the technique in Section 2, we quickly look at the related developments in searching document images.

There are two broad categories of techniques for search in document images. In the first category, search is carried out by converting the image into text using an OCR. This is often called recognition based approach. Second category does it mostly in the image domain by matching them in some appropriate feature space. This is referred to as recognition free approach. In this paper, we try to combine these two approaches.

Recognition based approaches are very popular, and easily scalable. Here text generated from OCR is used for indexing and retrieval. Much of the success of this approach lies in the performance of the OCR, but unfortunately this is not feasible for many of the non-Latin languages. Even for Latin languages, where robust commercial OCRs are present, there can be significant OCR errors for heavily degraded pages. Taghva [19], conducted a detailed study of how OCR errors effect the performance of retrieval based on a vector space model. In [20], Walker *et al.* have evaluated the effects of OCR errors on document clustering and topic modeling algorithms such as Latent Dirichlet Allocation (LDA). There have been many attempts to improve the quality of OCR output before indexing the text. These include post-processing of OCR texts using a lexicon or a dictionary of a particular language. There are also some methods such as [6, 8, 21] which reduce the lexicon size of a document with the knowledge of topic categorization. In [8], authors use a topic modeling framework for indexing OCR documents.

Due to the unavailability of robust OCRs, Indian scripts often resorted to the recognition free methods [3, 4, 13, 17]. Most of these originated from the idea of word spotting [15] where word images are represented using some features and comparison is done with the help of an appropriate distance metric. In traditional word spotting, query image is compared with every image in the database using a dynamic programming based matching. This makes it practically infeasible for large database, where Millions or Billions of word images are present. There have been attempts to speed it up with methods for approximate matching using LSH [10] and DBH [7]. Recently, BoVW based search shows a promising direction for complex datasets in Indian scripts [17, 22]. In BoVW, image patches are represented with the help of quantized features, known as visual words and the word image is represented as a histogram of visual words. Both [17] and [22] follow a very similar approach. In this work, we take this approach forward and combine with the noisy text available from the OCRs.

2. BASIC MODULES OF THE SEARCH SOLUTION

In this section, we propose a framework towards building an accurate large scale word image retrieval on the DLI books. We discuss the quantitative results in Section 3.2 on DLI. In this section, we discuss the design choices. We take the help of experimental verification of various ideas on a collection of four books each in Hindi and Telugu. Hindi is the national language of India which is used by more than 450 Million people across the world and Telugu is the regional language of Andhra Pradesh. It also has a strong presence of more than 75 Million speakers across the world. Table 1 shows the annotated dataset prepared using techniques discussed in [9] along with its ground truth. Here HP1 & TP1 are the dataset names given for Hindi and Telugu respectively. In all our methods, we evaluate using the standard Information Retrieval (IR) evaluation measures [12] such as the mAP and mPrec@10. The mAP is the mean area under the precision and recall curves for all the queries. While mPrec@10 shows mean of how many relevant results are obtained in top 10. Often user demands high precision.

Language	#Books	#Pages	#Words	Char %	Word %
Hindi(HP1)	4	420	112411	81.97%	54.83%
Telugu(TP1)	4	632	161274	79.48%	32.99%

Table 1: Dataset Details & OCR Accuracy

2.1 Optical Character Recognition

A typical OCR process consists of binarization of scanned image followed by applying some pre-processing steps such as skew correction, noise removal and image-text separation etc. The pre-processed image is segmented into lines, words and characters. The segmented character undergoes a feature extraction step and is given to a suitable classifier for recognition. Due to the complexity of Indian scripts, there are many challenges in each of these steps and none of them are trivial tasks. Some of the challenges with Indian scripts are (1) the presence of head line or "Shirorekha" in Hindi which results into difficulty in segmenting the characters properly (2) the number of unique classes (characters) are high which makes the classifier design challenging (3) vowel modifiers in conjunction with consonants make the character segmentation difficult. A detailed study of challenges in developing OCR for Indian languages is given in [2]. Because of these problems, the OCRs available for Indian languages are not robust enough to provide good search results based on the user queries. In Table 1, we show the performance of the state-of-the-art OCRs [2] available for Hindi and Telugu languages. We have consolidated the character and word level accuracies of each language used in this paper. It can be observed that the word level results are significantly low which make it difficult for a search engine to operate efficiently. The situation in DLI is far worse as one can see in

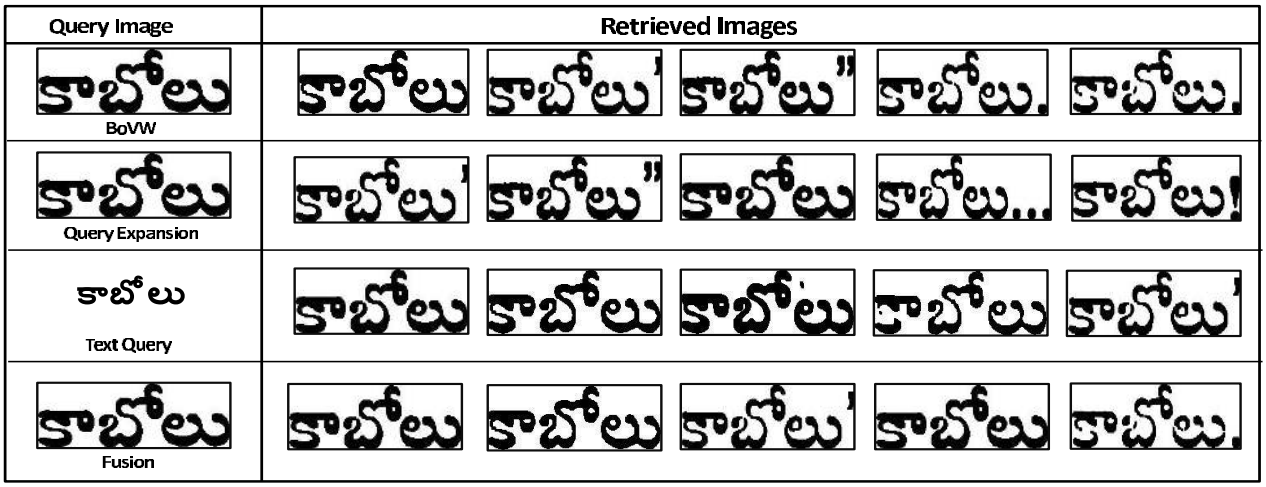

Figure 2: Sample retrieved word images based on different methods, which have same rank. Note that Text Query and Fusion give all clean results on the top. It happens because, for Text Query cleaned characters are used for visual word learning and Fusion OCR correctly recognises clean words.

Section 3. The working of search engine depends on the correctness of the inverted index which it creates by processing the text documents present in the corpus. Inverted index is also referred as posting file which holds the list of all unique words and records the number of occurrences of each word in each document. When a user posts a query, the search engine tokenizes it and searches for each token inside the inverted index. In Section 3.4, we discuss a page retrieval system developed on similar lines.

2.2 Bag of Visual Words

Bag of Visual Words (BoVW) representation is inspired from Bag of Words (BoW) representation in text retrieval. In BoW model, each document is represented by an unordered set of non-distinctive words present in the document, regardless of the grammar and word order. Document is formally represented with the help of frequency of occurrences (histogram) of the words in the vocabulary. These histograms are then used to perform document classification and retrieval. Analogously, an image is represented by an unordered set of non-distinctive discrete visual features. The set of these discrete visual features is called vocabulary. This visual vocabulary is then used to quantize the extracted features by simply assigning the label of the closest cluster centroid. The final representation for an image is the frequency counts or histogram of the quantized features $[h_1, ..., h_k]$ where h_i is the number of occurrences of i^{th} visual word in the image and k (determined empirically) is the vocabulary size. To account for the difference in the number of interest points between images (due to size etc.), the BoVW histogram is normalized to have unit L1 norm.

For interest point detection, we use Harris corner detector which is proven useful in image representations [16]. At each of these interest points, we extract a Scale Invariant Feature Transform (SIFT) [11] descriptors to describe the local information as a vector of gradients. Vocabulary creation is done using computationally efficient Hierarchical K Means (HKM) [14].

Baseline results of OCR and BoVW based search engine are shown in Table 2. It can be observed that, mAP of BoVW based retrieval is improved by 20% for Hindi and 40% for Telugu as compared to OCR. Even in top-10 results, accuracy rate of OCR is lower compared to BoVW.

Dataset	#Query	OCR Search		BoVW Search	
		mAP	mPrec@10	mAP	mPrec@10
HP1	100	42.64	69.90	62.54	81.3
TP1	100	29.64	44.30	71.13	78

Table 2: OCR and BoVW Baseline Results

2.3 Query Expansion

Query Expansion (QE) is simple and efficient method for improving search performance in text search engines. In QE, highly ranked retrieved result is used to formulate the query again and better results are retrieved based on new improved query. If top ranked selected results are mostly correct, then the new formulated query will have more relevant information. Chum *et al.* [5] first introduced automatic query expansion in visual domain. The popular and efficient one is average query expansion. In average query expansion method, for a given query image, BoVW vectors of top results are averaged together with the BoVW vector of query, and this resulting query expanded BoVW vector is used to re-query the index. We have also used similar approach for query expansion. We take top-p queries from the retrieved list and formulate the new query based on their visual words. While assigning weight to each of these visual words, we have taken rank of the word also into the consideration. Weight factor (determined empirically) is defined as $1/2^{rank}$. We also observed that few of the visual words are present in all of the considered top results. This implies that these visual words are specific to given query words. Therefore, we have increased weight of these visual words by two fold. Results of QE are shown in Table 3. As it can be observed, QE improves 2% in terms of both mAP and mPrec@10 as compared to BoVW which is shown in Table 2.

Dataset	#Query	Query Expansion		Text Query		Naive Fusion		Eds-Fusion		Hybrid-Fusion	
		mAP	mPrec@10	mAP	mPrec@10	mAP	mPrec@10	mAP	mPrec@10	mAP	mPrec@10
HP1	100	66.09	83.86	56.32	73.89	75.66	90.7	79.58	90.8	80.37	91.4
TP1	100	73.08	79.89	69.06	78.83	76.02	81.2	78.01	81.4	80.23	83.7

Table 3: Performance Statistics with Enhancements and Fusion Techniques

2.4 Text Query Support

One of the eventual goals of document image retrieval community researcher is to support text query in document image retrieval system. Here we propose a novel and simple framework for text query support. In BoVW approach, all images are represented as histograms. Indexing and retrieval are done based on these histograms. We observed that the histograms of all variants of a given character are more or less similar. Only difference arises for some of cluster centres where their frequencies may vary, but cluster centres will be same.

Based on the above observations, we have selected a small subset data with ground truth such that all possible characters are present atleast twice in it and learnt the character specific visual words. First of all, these selected words are segmented at character level using connected component and visual words of each character are found using corresponding ground truth. For a given character with different variations, we have two or more sets of visual words. To take care of variations, we take average of visual words of all variants of that character. Thus given a textual query, we can synthesise the BoVW histogram of that word. Since our histogram is obtained by composing character histogram, "context" is missing in many SIFT vectors. Due to this, recall for this retrieve list is not high but initial precision is very high. To improve recall, we have also indexed word images and used query expansion as explained in Section 2.3 to formulate query histogram based on initial results of text query index. Results are shown in Table 3 and we can observe that text query results are comparable to word image.

2.5 Fusing OCR results with BoVW

We further improve the search on document images for languages which do not have highly accurate OCR systems. There are situations where a word is identified properly by BoVW approach but OCR has produced a wrong word. The reverse condition is also applicable. We observe that the OCR system has got a very high precision while BoVW based algorithm has got a very high recall. In other words, for a given query, the words returned by an OCR system mostly matches to the actual query word unless there is high amount of post-processing. But we cannot assure a high recall with OCR results since the languages that we deal here do not have robust OCR systems over the corpus used in this paper. In the case of BoVW, it associates a cost with all word images. The results are sorted in decreasing order of scores and filtered above a particular threshold. This gives BoVW much higher recall than OCR results. Using these assumptions, we propose different methods in which we fuse the results obtained using OCR and BoVW based search. In our experiments we always keep the OCR results on top of search results due to its high precision and followed by BoVW. The order in which the BoVW results are given plays the most important role for deciding the mAP and mPrec@10. In our fusion system, it is important to have the alignment between the BoVW and OCR sources to avoid duplicates in final search results. We use the word coordinates of OCR and BoVW algorithms and align them by finding the maximum intersection area.

A simplest way to think of a fusion is to concatenate the search results of OCR text search followed by BoVW results. We call this naive fusion. In Table 3, we show the performance statistics of various fusion methods which are present in this paper. As expected, the mAP and mPrec@10 values have increased as compared to the baseline results which are shown in Table 2. The precision values increases since there are some words retrieved by OCR which are not present in BoVW output list. The same reason also improves the measure of mPrec@10 as we have placed the OCR results at the beginning of the fusion results. The mAP of fusion results for clean datasets HP1 and TP1, has increased around 9% while mPrec@10 for Hindi is increased by around 7% and for Telugu mAP is increased by 3% and mPrec@10 by 2% as compared with query expansion results.

In order to improve on the naive method, we take help of two different information. The first one is the score given by retrieval system for each retrieved word image in BoVW. The second one is in the use of edit distance (EDS) and modified edit distance (MEDS) values for a given query. The edit distance or the Levenshtein distance between two strings (string1 and string2) tells the costs of insertion, deletion and substitution for modifying a string1 into string2. In a normal edit distance which we call as EDS, the costs given for insertion, deletion and substitution are equal while in the case of MEDS, the substitution cost is modified [8] as shown in Equation 1.

$$Sub\_Cost(char1, char2) = 1 - Conf(char1, char2) \quad (1)$$

Here, Sub_Cost is the substitution cost given to the characters classes $char1$ and $char2$ and $Conf$ is OCR confusion matrix. It represents the probability of misclassification at the symbol level by listing down the probability of assigning a symbol with itself and all other class symbols of that language. In edit distance based fusion (Eds-Fusion), we modify the ranked order of results obtained by BoVW algorithm and concatenate it with OCR results. Since we have done alignment between OCR and BoVW, we find out the OCR texts for each word images returned by BoVW. We then calculate the EDS and MEDS values for that OCR text with the query and apply rules to improve the order of BoVW results. We compare the results of Eds-Fusion with naive fusion as shown in Table 3. mAP gains a 4% of improvement in Hindi and 2% for Telugu.

We observed that different rank order lists are obtained when different images of the same word are taken as queries. We formulate these different queries from OCR output. Ranked

aggregation methodology is used to combine obtained different rank order lists. In this method, more weight is given to retrieved word images in top of both the lists. We call this fusion as Hybrid fusion. The combined score is defined as follows:

$$finalScore = \Sigma_{i=1}^{n} Score_i \times (1/Pos_i) \quad (2)$$

where $Score_i$ is score from list i and Pos_i is corresponding position.

2.6 Discussions

In this section, we discuss some of the cases where BoVW is able to recognize a word image but OCR fails to do so and vice-versa. In Figure 3, (a) & (c) show the actual query word images and (b) & (d) show one of the top results obtained from BoVW algorithm respectively. For a BoVW algorithm the results are very much similar in terms of its visual representation but actually its meaning is very different in the text domain. These sort of errors occur in BoVW irrespective of a clean word image but a similar situation is less likely to create an error in OCR. It is known that an OCR is severely affected by degradation but for many cases similar to the image that we have shown, it works perfectly.

Figure 3: Failure cases of BoVW where OCR succeeds.

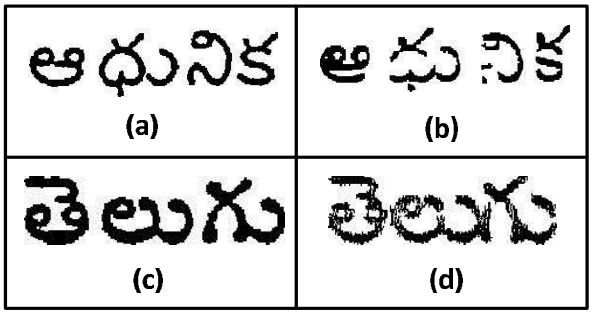

Figure 4: Failure cases of OCR where BoVW succeeds.

There are also situations where OCR fails but BoVW still succeeds. In Figure 4, (a) & (b) show a word image taken from TS1 and TP1 datasets for the same query respectively. We can see that the image shown in (b) is degraded due to cuts which makes it difficult for an OCR to give correct output. In Figure 4(d), we have shown a different degradation of merges where again an OCR fails. It is observed from our experimentation that BoVW algorithm can withstand to some level of degradation. However extending this work to severely degraded cases is a part of future task.

3. SEARCHING IN DLI IMAGES

In this section, we discuss the details of our search experiments on DLI pages.

3.1 Dataset & Baseline Results

Language	Ann.	#Books	#Pages	#Words	Char %	Word %
Hindi(HS1)	Yes	11	1000	362593	77.14%	26.10%
Hindi(HS2)	No	52	10196	4290864	-	-
Telugu(TS1)	Yes	11	1000	161276	80.75%	15.69%
Telugu(TS2)	No	69	13871	2531069	-	-

Table 4: DLI Dataset Details & OCR Accuracy

DLI hosts a very large scale dataset for testing and evaluating various strategies for OCRs and word image retrieval methods. Table 4 shows the subset of the DLI corpus that we have used in this work. It contains around 10K pages and 4.3M words for Hindi and around 14K pages and 2.5M words for Telugu. We have divided the entire dataset into two sets. Set1 contains around 1000 pages of annotated data prepared using [9] along with its ground truth. This is primarily used for evaluating the mAP and mPrec@10 and validate the efficiency of the proposed method in comparison with the ground truth. We have named Set1 datasets as HS1 and TS1 for Hindi and Telugu respectively. Set2 is the larger set which corresponds to the un-annotated pages where we don't have ground truth and hence cannot calculate mAP on it. In this set, we could only employ mPrec@N measure and the results are manually validated to check its correctness. Similar to Set1, we have named Set2 datasets as HS2 and TS2 respectively. In the same table, we have shown the character and word level accuracies of the existing OCR systems. Since DLI dataset is more challenging than the dataset used in Section 2, the accuracies are much lower which further decreases the efficiency of a search engine running on top of it.

Figure 5 shows some sample word images from the books present in the DLI corpus. As one can clearly see that the images have severe degradation due to cuts, merges, bad print quality and salt and pepper noise, etc. In our collection, we also have 18th century books whose fonts are not in use any more. There are also cases where some of the characters exist in their older form of writings. Under these situations, the performance of the current state-of-art OCR systems for these languages degrade even more.

3.2 Quantitative Evaluation

In this section, we show the results obtained by applying the same techniques discussed in the previous section on a much larger DLI dataset which is more challenging in terms of its size and quality. In Table 5, the third and fourth columns show the baseline results for OCR and BoVW algorithms on the selected queries respectively. The mAP of OCR is about 14.95% for Hindi and 27.03% for Telugu. Considering the case of BoVW, it is almost consistent with Tel-

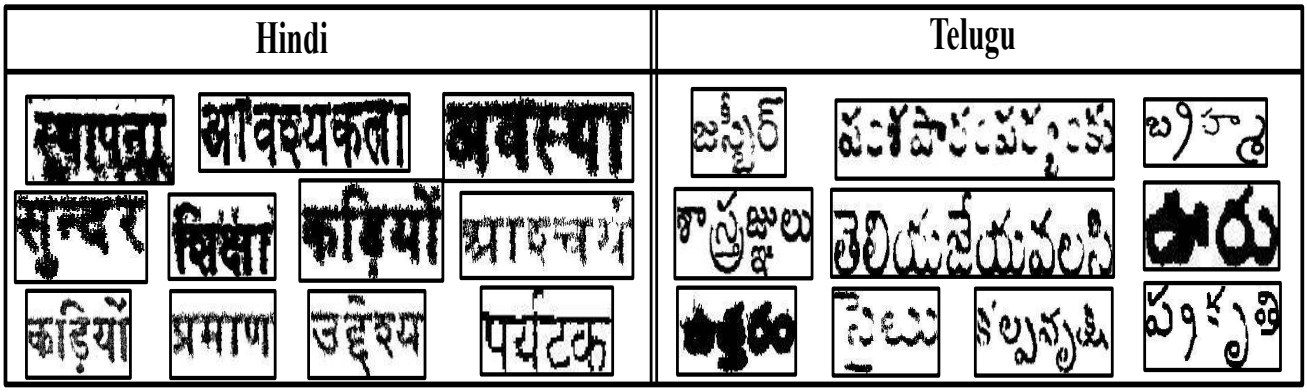

Figure 5: Sample Word Images from the Corpus. Observe Degradations.

Dataset	#Query	OCR		BoVW		Fusion	
		mAP	mPrec@10	mAP	mPrec@10	mAP	mPrec@10
HS1	100	14.95	62.60	60.55	95.5	68.81	95.6
TS1	100	27.03	62.10	74.38	90.6	78.41	91.9

Table 5: Performance Statistics on DLI Annotated Dataset

ugu performing better than Hindi. The last column shows the results after applying the fusion technique. The mAP has increased about 8% for Hindi and 4% for Telugu. Similarly we can see that mPrec@10 values have increased to around 95% for Hindi and 92% for Telugu. It means that the top order results are mostly correct after doing fusion. Table 6 shows the test conducted on the larger un-annotated datasets HS2 and TS2. We have selected around 50 queries each and manually verified the mPrec@N (N = 10, 20 and 30). The reason of measuring precision at three different levels is because the dataset is huge and the number of occurrences could be more for each query and mPrec@10 would be high for all methods. mPrec@N values for these three levels are more that 93% for both the languages.

In Figure 7, we show the fusion results for both languages along with their query images. We can observe that the proposed system is able to retrieve word images having severe degradation and print variation. It is also able to handle cut and merges which are the most common forms of degradations in document images.

Dataset	#Query	Precision@N	OCR	BoVW	Fusion
HS2	50	mPrec@10	82.03	96.94	97.11
		mPrec@20	75.16	94.83	95.42
		mPrec@30	71.12	92.82	93.16
TS2	50	mPrec@10	90.85	99.14	99.14
		mPrec@20	85.42	98.00	98.85
		mPrec@30	80.76	96.38	96.57

Table 6: Performance Statistics on DLI Unannotated Dataset

We have used Lucene [1] which is a popular, reliable and open source search engine for all the indexing purposes. Lucene indexes documents using inverted file index consisting of multiple sub-indexes. One advantage of these sub-indexes is that it can be searched independently and it maintains almost constant searching time even index size increases. We have performed all our experimentation on a system with 2 GB RAM and Intel® Core TM 2 Duo CPU with 2.93 GHz processor. Set1 dataset has index size as 800MB and retrieval time as 500ms and Set2 dataset has index size as 6GB and retrieval time as 900ms.

[1]Lucene: http://lucene.apache.org/.

3.3 Analysis

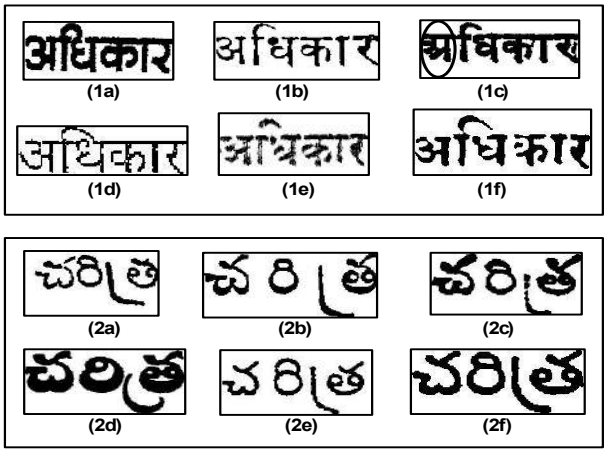

Figure 6: Script and Font variations for a Query Image across the Dataset: Figure (1a)-(1f) shows variations in a query image for Hindi language and Figure (2a)-(2f) shows variations for a Telugu keyword.

Script & Font Variations

The dataset that we have used contains books published from year 1624 and as one would expect there would be many changes in the script, font and typesetting over the years. An accurate word image retrieval on such a dataset is a real challenge. Figure 6 shows one sample query taken from each language and its variation in terms of font and script

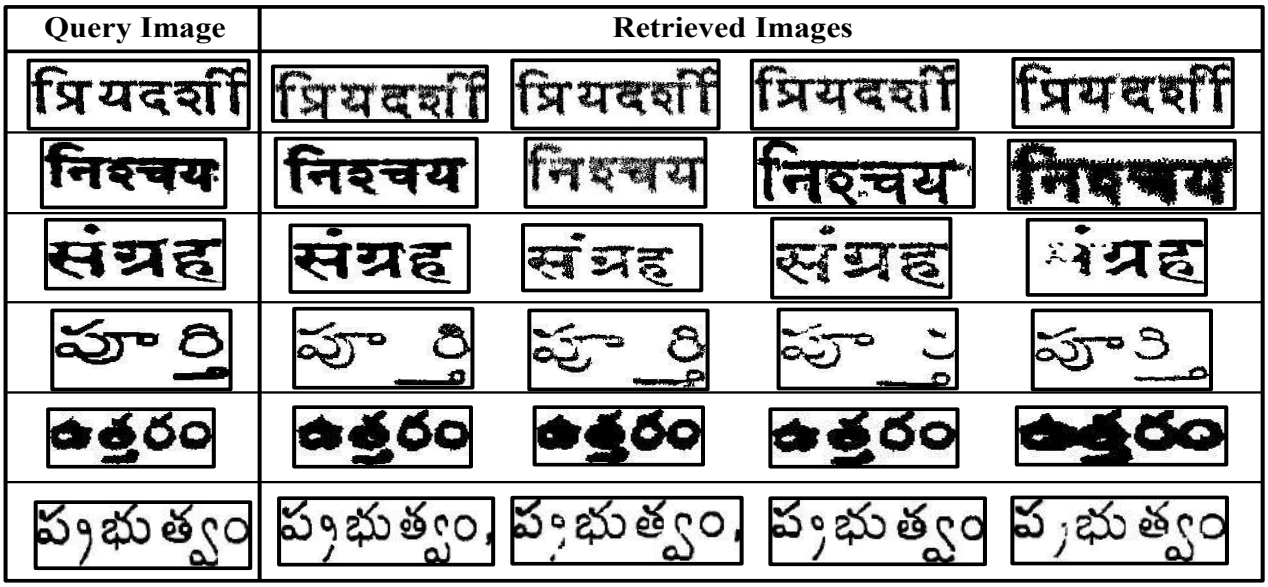

Figure 7: Retrieved Results on Datasets: First Column shows Query Images. Second Column Shows the Retrieved Results in decreasing order of Rank.

across the dataset. The first symbol of Figure 6(1c) shown in an ellipse is an older way of writing the symbol which is not used these days. For such a case, an OCR will fail to recognize since it is trained only on a particular character set. However in the case of BoVW, it has the ability to capture the contextual information of a word. Hence in the case of Figure 6(1c), BoVW will still retrieve the image using the context of other characters.

mAP vs Word Length

The mAP associated with BoVW is dependent on the length of the query. For longer queries, mAP increases since the discriminative power is more with the increase in number of visual words. But in the case of OCR, the probability of word error increases with the increase in number of characters present in a word. Hence fusing these results gives more mAP than the individual methods.

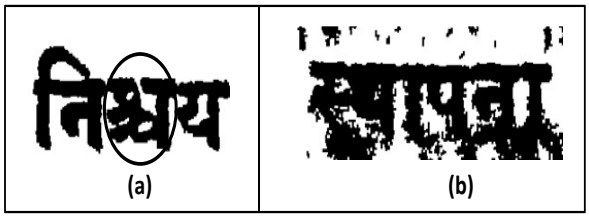

Figure 8: Failure cases for both BoVW and OCR due to old script and lots of degradation.

Failure Cases

The failure cases of the fusion method are those where both OCR and BoVW fail. In Figure 8(a), we show the word image which is smaller in length and contains a character shown in ellipse, which is not used nowadays. An OCR will fail to recognize that word properly if it is not trained on such a character. Since it is a smaller word image, BoVW is not able to capture much of the context and will fail to show the word image results if similar character is not there in query image. The next example shown in Figure 8(b) is a highly degraded image which also didn't come up in the fusion system. It failed in both BoVW and OCR.

3.4 Page Level Retrieval

We have modified our word level retrieval to page level retrieval. For page level retrieval, we have indexed page level details. When two or more words from same page are retrieved in top we have combined their corresponding scores and ranked the list based on page scores. Figure 9 shows a Hindi page retrieved from HS2 dataset based on query as highlighted in the figure. The page which is shown in the figure contains 6 occurrences of the query word and is retrieved on top from collection of all books. In page level retrieval, it is always important to get pages having large number of occurrences of query word on the top. This is achieved in our word level system without any extra cost.

3.5 Discussions

In this section, we try to anticipate the possibilities to take the fusion to the next level.

Learn to improve from annotated dataset

In Section 2.5, we explained the use of confusion matrix to deal with the errors given by OCR. The same concept can be inherited to a BoVW domain where we find out the visual confusion words for the algorithm. We can use the annotated datasets HS1, TS1 present with its ground truth for improving the results.

Necessity of Costly Features for Re-ranking

We have observed from our experiments that some images, see Figure 8, are hard to retrieve without post processing of

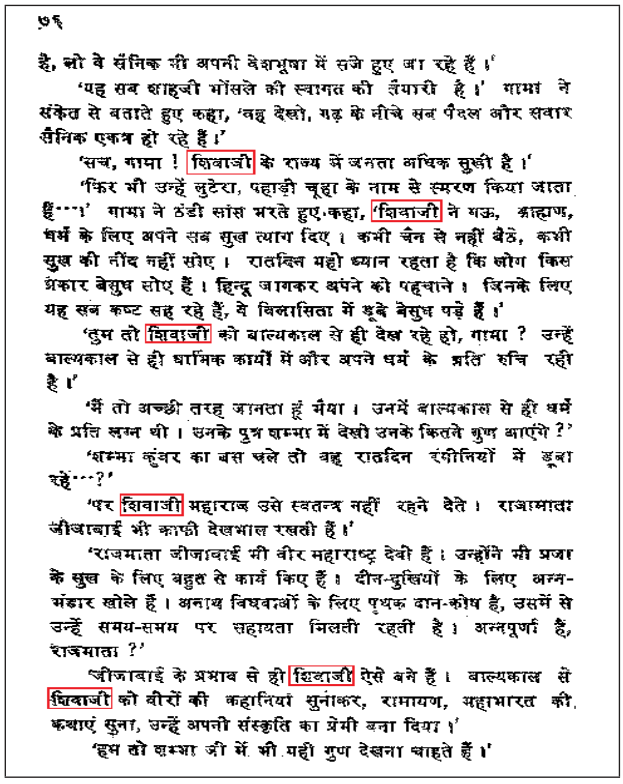

Figure 9: Sample Retrieved Page from Book name *Maharastra* Published in 1927 having Historical significance. Here searched word in '*SHIVAJI*', the founder of Maratha Empire and all 6 occurrences of the query word are correctly identified and marked in the image.

retrieved ranked list and need extra effort in bringing them to top of the rank list. One of the possibilities is to use some costly features for re-ranking. This will increase the time but improve the performance. For huge dataset like DLI, this can be done after displaying initial results to the user and refine the result when user demands more results.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a document retrieval system by fusing BoVW and OCR. We have provided a text based search engine. The efficiency of the method is shown experimentally on DLI dataset of two Indian languages. We have demonstrated system performance on more than 100 Books and reported a mean average precision of around 0.8 and mean precision of more than 0.9. We are also able to demonstrate page level retrieval. In future, we would like to provide semantics in search and fuse multiple features online.

5. ACKNOWLEDGMENT

This work was partly supported by Ministry of Communication and Information Technology, Government of India.

6. REFERENCES

[1] V. Ambati, N.Balakrishnan, R. Reddy, L. Pratha, and C. V. Jawahar. The Digital Library of India Project: Process, Policies and Architecture. In *ICDL*, 2007.

[2] D. Arya, C. Jawahar, C. Bhagvati, T. Patnaik, B. Chaudhuri, G. Lehal, S. Chaudhury, and A. Ramakrishna. Experiences of integration and performance testing of multilingual OCR for printed Indian scripts. In *JMOCR and AND*, 2011.

[3] A. Bhardwaj, S. Kompalli, S. Setlur, and V. Govindaraju. An OCR based approach for word spotting in Devanagari documents. In *DRR*, 2008.

[4] S. Chaudhury, G. Sethi, A. Vyas, and G. Harit. Devising Interactive Access Techniques for Indian Language Document Images. In *ICDAR*, 2003.

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In *ICCV*, 2007.

[6] F. Farooq, A. Bhardwaj, and V. Govindaraju. Using topic models for OCR correction. *IJDAR*, 2009.

[7] E. Hassan, S. Chaudhury, and M. Gopal. Word shape descriptor-based document image indexing: a new DBH-based approach. *IJDAR*, 2012.

[8] E. Hassan, V. Garg, S. Haque, S. Chaudhury, and M. Gopal. Searching OCR'ed Text: An LDA Based Approach. In *ICDAR*, 2011.

[9] C. V. Jawahar and A. Kumar. Content-level Annotation of Large Collection of Printed Document Images. In *ICDAR*, 2007.

[10] A. Kumar, C. V. Jawahar, and R. Manmatha. Efficient Search in Document Image Collections. In *ACCV*, 2007.

[11] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] M. Meshesha and C. V. Jawahar. Matching word images for content-based retrieval from printed document images. *IJDAR*, 2008.

[14] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *CVPR*, 2006.

[15] T. M. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 2007.

[16] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of Interest Point Detectors. *IJCV*, 2000.

[17] R. Shekhar and C. V. Jawahar. Word Image Retrieval Using Bag of Visual Words. In *DAS*, 2012.

[18] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[19] K. Taghva, J. Borsack, and A. Condit. Effects of OCR errors on ranking and feedback using the vector space model. In *IJIPM*, 1996.

[20] D. Walker, W. B. Lund, and E. K. Ringger. Evaluating Models of Latent Document Semantics in the Presence of OCR Errors. In *ICEMNLP*, 2010.

[21] M. Wick, M. Ross, and E. Learned-Miller. Context-Sensitive Error Correction: Using Topic Models to Improve OCR. In *ICDAR*, 2007.

[22] I. Z. Yalniz and R. Manmatha. An Efficient Framework for Searching Text in Noisy Document Images. In *DAS*, 2012.