Reverse Annotation based Retrieval from Large Document Image Collections

Pramod Sankar K.
Center for Visual Information Technology, IIIT-Hyderabad, India
pramod_sankar@research.iiit.ac.in

ABSTRACT

A number of projects are dedicated to creating digital libraries from scanned books, such as Google Books, UDL, Digital Library of India (DLI), etc. The ability to search in the content of document images is essential for the usability and popularity of these DLs. In this work, we aim toward building a retrieval system over 120K document images coming from 1000 scanned books of Telugu literature. This is a very hard problem because: i) OCRs are not robust enough for Indian languages, especially the Telugu script, ii) the document images contain large number of degradations and artifacts, iii) scalability to large collections is hard. Moreover, users expect that the search system accept text-based queries and retrieve relevant results in interactive times.

We propose a *Reverse Annotation* framework [1], that labels word-images by their equivalent text label in the offline phase. Reverse Annotation applies a retrieval based approach to recognition. It first identifies a set of keywords which are considered useful for labeling and retrieval. Exemplars are obtained for each word from a crude OCR or human annotations. The labels are then propagated across the rest of the collection by matching words in the image-feature space. Since such a matching is computationally expensive, scalability is achieved using a fast approximate nearest neighbor technique based on Hierarchical K-Means. Our framework allows us to assign text labels for document images offline, allowing us to build a search index for quick online retrieval. An example query and the retrieved results are shown in Figure 1. We are unaware of any conventional OCRs which can recognize such images.

There are three major contributions of our work: i) recognizing the entire document collection together, instead of one-at-a-time, ii) speeding up recognition by clustering multiple instances of a given word, iii) recognising at the word-level, avoiding the pitfalls of character segmentation and recognition.

Using the techniques developed from my work, we were able to successfully build a retrieval system over our challenging dataset. To the best of our knowledge, this is the largest collection of document images that have been made searchable for any Indian language. Our algorithm is easily scalable to larger collections, and directly applicable to documents from other language scripts.

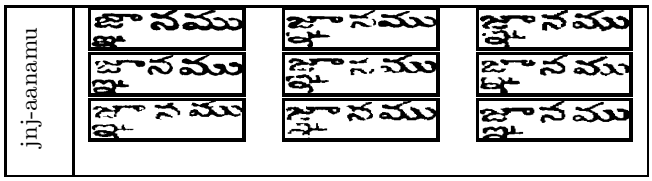

Figure 1: The retrieved word-images for an example query. These words are hard to recognize with a conventional OCR. Inspite of heavy degradations, our framework correctly retrieves document-images containing the query.

The first issue to discuss, is the fraction of word-images that remain unrecognized at the end of the Reverse Annotation phase. Words of the vocabulary that are not very popular, or are not contained in the training data are not labeled for in the test set. It is imperative that we estimate the cost of not being able to answer such queries. If this cost is indeed high, we need to explore methods to label such infrequently occurring words in the collection. Needless to say, such methods should be computationally efficient without compromising on accuracy.

The other major issue to discuss is the evaluation of retrieval results. The true recall of the retrieval system cannot be computed, since it is impossible to identify every occurrence of the given query in such large data. Is Precision alone a sufficient indicator of the retrieval performance. Moreover, the precision coincides with the accuracy of the underlying annotation module. How would one evaluate the retrieval performance at a document level, that is not the same as recognition accuracy. Finally, how could one quickly estimate the satisfaction of a user's information need.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Miscellaneous

Keywords

Document Images, Recognition-free, Scalability

1. REFERENCES

[1] Pramod Sankar, K. and C. V. Jawahar. Probabilistic reverse annotation for large scale image retrieval. In *Proc. CVPR*, 2007.