Recognition of Books by Verification and Retraining

Neeba N.V. and C.V. Jawahar

Recognition of Books is different from recognition of isolated document images.

Traditional OCRs.

- Designed for isolated pages.
- Performance remains same over time – the classifier repeats the same mistake.
- Training is offline (apriori done) – no scope for improvement in performance.

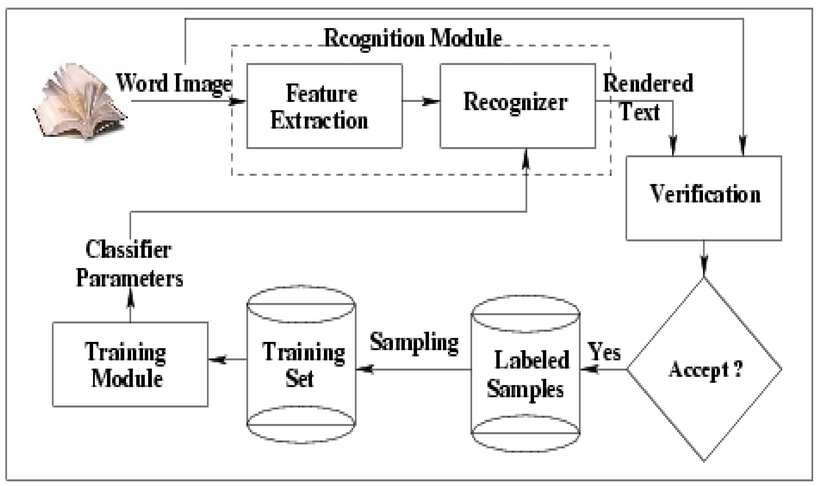
Role of Books.

- Digital libraries need the recognition of complete book.
- Books are typeset in the same font and style.
- We can use the samples from first few pages to learn the classifier, and obtain better performance over the rest of the collection.

Book Recognizer.

- Designed specifically for books/ large collection of documents.
- Adapt the classifier to the font and style of the collection and improve the performance over time.
- *New* training data is introduced into the system without any manual intervention.

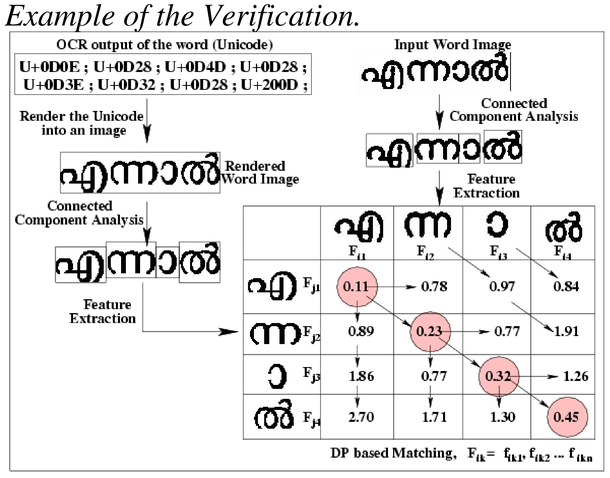
Overall Architecture of the Book Recognizer .



We employ a data-driven adaptation method to enhance the performance of the classifier specific to a particular collection.

- Our focus is limited only to the recognizer / classifier, (and not to the pre-processing stages).
- An automatic learning framework to improve the performance of the classifier over iteration.
- This is achieved with the help of a high performance *verification module*, which acts as a post-processor in the system.
- The *Verification module* collects / verifies the samples from the book, labels and stores them.
- The new samples from the labeled set are added to the training set and the classifier is retrained online to create an improved classifier. The new classifier is used for further classification.
- This process repeats until the improvement in the performance is less than a threshold.

An Example of the Verification.

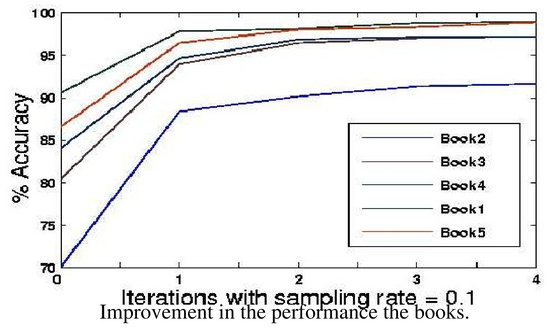


- Input : Word Image + text/ Unicode from OCR.
- Output : Matched symbols/ characters (+ information related to cuts and merges etc).
- Procedure : A dynamic programming (DP) based matching of sequence of feature vectors.

Experiments and Results

Book No.	*# Pages*	*# Words*	*# Symbols*
Book 1	96	11,404	74,774
Book 2	119	20,298	147,652
Book 3	84	10,585	83,914
Book 4	175	21,292	152,204
Book 5	94	12,111	92,538

Details of the books used for the experiments.



Improvement in the performance the books.

Iteration	*0.01*	*0.05*	*0.1*	*0.3*
0	80.42	80.42	80.42	80.42
1	93.28	94.33	93.96	93.94
2	96.46	96.80	96.47	95.35
3	97.41	97.59	97.01	96.28
4	97.52	97.69	97.26	96.46

Accuracies obtained with varying sampling rate for Book3.

The experiments are conducted on 500000 samples from five books in Malayalam (an Indian language).

We obtain an average performance improvement of around 14% in classification accuracies.

The change in learning rate with the change in the sampling rate is marginal.

Center for Visual Information Technology
International Institute of Information Technology Hyderabad
http://cvit.iiit.ac.in