CLEAT:
A CLassification, Enhancement and Analysis Toolkit for Heterogeneous Document Image Collections

Phase I Report

Centre for Strategic Infocomm Technologies
Singapore

from

Laboratory for Language and Media Processing
University of Maryland, College Park, MD, USA

September 20, 2006

Table of Contents

1 Summary

In this phase, we focused on data collection, refining algorithms for DocLib and enhancements to our ground truthing interface. There were several staff changes on our programming staff and the changes that we anticipated being completed in April were not completed until recently. In this report , we have organized the content slightly differently then the original deliverables.

First, we provided a CD of data during the CSIT visit, and have augmented it with additional data that will be provided electronically. Details of the collection are provided in Section 2 including what metadata is represented, what type so of ground truth have been added and the distribution of the dataset.

Second, we describe our evaluation framework in Section 3, including the modifications that have been made to the ground truthing interface to support evaluations

Third we provide discussion of several module enhancements made in Phase I.

Phase 1 - Deliverables

- Deliver completed CLEAT data collection.
- Provide ground truth for subset of data including signatures, stamps, logos, handwritten, and machine printed text.
- Provide document describing evaluation framework.

2 CLEAT Data Collection

The CLEAT data collection now consists of almost 20,000 document images between the classification dataset (6620) and the remainder of purely document images. Our goal is to provide a single dataset with metadata contained in an XML file describing each file.

2.1 MetaData

The Metadata is provided in a GEDI/DocLib supported metadata format which has been widly used in our development For this project, we have added the following page level attributes:

- Image Class: Document, Image or Image with Text
- Page Type: Handwitten, MachinePrint, Mixed
- Primary Language:
- Genre-Class: Drawing, Form, Article, etc

Each image in this collection has an associated XML file. The representation is extensible and adding attributes at the page or zone level will not effect the use of our tools such as GEDI. Here is a sample of the header information for a document of handwritten Korean.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--GEDI is developed at Language and Media Processing Laboratory, University of
 Maryland.-->
<GEDI xmlns="http://lamp.cfar.umd.edu/GEDI" version="1.0">
    <DL_DOCUMENT src="KOR0430001.tif" NrOfPages="1" docTag="xml">
```

```
                    <DL_PAGE gedi_type="DL_PAGE"
                                  src="KOR0430001.tif"
                                  pageID="1" width="2592" height="3300"
                                  Quality="Good"
                                  ImageClass="Document"
                                  PageType="Handwritten"
                                  PrimaryLanguage="Korean"
                                  Source="ScannedMedia"
                                  GenreClass="Notes"
                                  Misc="Blank">
                    </DL_PAGE>
               </DL_DOCUMENT>
          </GEDI>
```

2.2 Ground Truth Subset

We have ground truthed a subset of the document collection for logos, signatures and stamps, and we have provided the zone level metadata in our collection. In addition, we have run automated algorithms on a section of the data to produce zone level results.

2.3 Dataset Distribution

The current distribution of our dataset is as follows.

Genre	
Forms, Drawing, Tables et at.	
Forms	644
Drawing	42
Tables	100
Chemistry formulae	25
Math equations	165
Figures	40
Total	1016
Business documents and Memo letters	
Business documents clean	52
Business documents degraded	2700
Business documents with annotations	160
Memo letters (English + Multilingual)	978
Total	3890
Journal and Conference Papers, Articles	
English	2785
German	359
Japanese	478
Total	3622
Newsletters and Flyers	
Google images	1417
Arabic Newswire + Broadcast News	338

Total	1755
Structured Documents	
Phonebook	229
Dictionaries (Chinese English, English Chinese)	1148
Yellowpage	84
Total	1461
Handwritten	
Arabic	60
Chinese	146
Cyrillic	410
Japanese	47
Korean	80
Thai	319
Hindi	281
Total	1343
Page Classification Datasets (Google Image)	
Document	757
Image with Text	2443
Non-Document	3420
Total	6620
Total in all genre categories	19707

3 Evaluation Protocol and Framework

We have completed and implemented an evaluation framework for this project. It consisted to changes to the GEDI tool to allow the visualization of results. The changes to the interface where more significant then first envisioned, and the interface now has "command" capabilities to run and visualize various DocLib modules. We have integrated two flavors of evaluation. One evaluation for detection modules and another for recognition and classification. In Phase I we focused on detection.

For example, for logo detection, we have

$$\text{Accuracy} = \frac{\text{\# of correctly detected logos}}{\text{\# of logos in groundtruth}}$$

$$\text{Precision} = \frac{\text{\# of correctly detected logos}}{\text{\# of detected logos}}$$

While the community has published a number of works related to evlaution, there is not single accepted standard, we hope that a public tool such as those being developed will change that trend. Here are some relavant research references:

(Belaïd and Pierron; Kanungo, Marton et al.; Liang, Phillips et al.; Patton and Patton 1987; Kanai, Rice et al. 1995; Trier and Taxt 1995; Blue, Candela et al. 1998; Junker and Hoch 1998; Junker, Hoch et al. 1999; Liang 1999; Mao and Kanungo 2001; Wang, Haralick et al. 2001)

Rather then explicitly describing the framework, we describe the approach in three sections below:

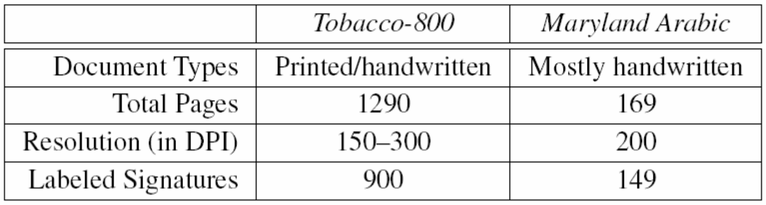
4 Evaluation of CLEAT Modules

4.1 Signature Detection

4.1.1 Datasets

To evaluate the structural saliency approach for signature detection on multiple languages, we used two large collections of real world documents—Tobacco-800 dataset and the University of Maryland Arabic dataset. Tobacco-800 is a public subset of the IIT CDIP Test Collection [1, 2], based on 42 million pages of documents (in 7 million multi-page TIFF images) obtained from UCSF [3] and released by tobacco companies under the Master Settlement Agreement. Tobacco-800 is a realistic dataset for document analysis and retrieval as these documents were collected and scanned using a wide variety of equipment over time. In addition, a significant percentage of Tobacco-800 are consecutively numbered multi-page business documents, making it a valuable testbed for various content-based document retrieval approaches. The Maryland Arabic dataset consists mainly of Arabic handwritten business documents. Using public datasets gives more realistic evaluation in contrast to common published evaluations using self collected datasets that captures much less variations. Typical dimensions of documents range from 1200 × 1600 to 2500 × 3200 pixels in Tobacco 800 and 1700 × 1200 to 2000 × 2600 pixels in Maryland Arabic.

Table 1. Summary of the English and Arabic evaluation datasets.



	Tobacco-800	*Maryland Arabic*
Document Types	Printed/handwritten	Mostly handwritten
Total Pages	1290	169
Resolution (in DPI)	150–300	200
Labeled Signatures	900	149

The groundtruth of signatures were manually labeled in rectangular boxes using our developed Java editor. Whenever possible, we also label the identity of the signer by reconciling the document context. This enables quantitative evaluation on signature retrieval, where the identities of the signers are required. Since the number of signatures varies significantly across documents, we assume no prior knowledge on the distribution of signatures per document. In our evaluation, we use all the documents in the Tobacco-800 and Maryland Arabic datasets.

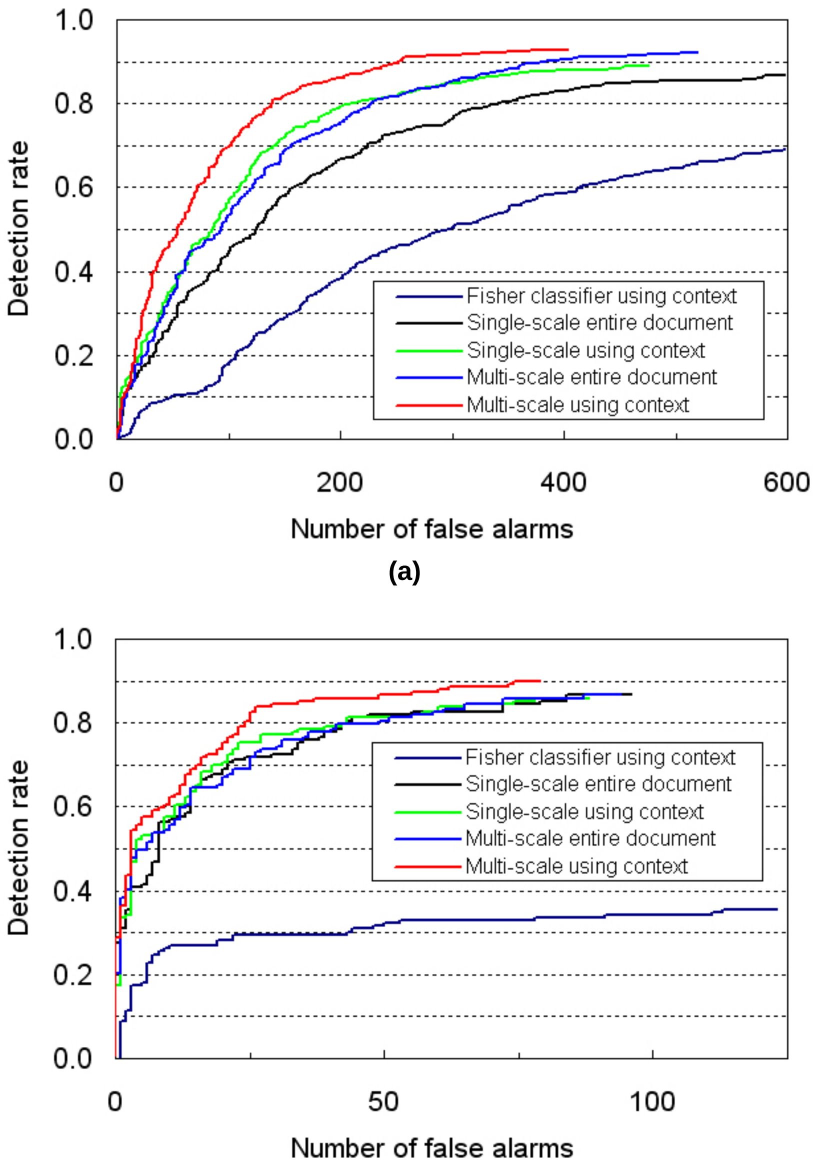
4.1.2 Evaluation Methodology

We focused on two aspects in our evaluation. First, we use the detection probability $\boldsymbol{P_D}$ and false-alarm probability $\boldsymbol{P_F}$ as metrics. P_D and P_F represent the two degree of freedom in a binary hypothesis test and they do not involve a prior probabilities of the hypothesis. To factor in the “quality” of the detection, we consider a signature correctly detected and complete if the detected region overlaps with more than 75% of the labeled signature region. We declare a false alarm if the detected region does not overlap with more than 25% of any labeled signature region.

4.1.3 Results and Discussion

Fig. 1 shows the ROC curves on the Tobacco-800 and Maryland Arabic datasets. Fisher classifier using size, aspect ratio, and spatial density features serve as a baseline for comparison, with all other procedures remaining the same in the comparison experiment. We use two scale levels in multi-scale detection experiments. Parameters involved in obtaining the ROC curves, including the cutoff threshold in saliency and estimated signature dimensions, are tuned on 10 documents. We use the following approach to compute each operating point on an ROC curve. After we compute the saliency of each signature candidate, we store it with the internal zone representation of the candidate. We apply a reasonably low global decision threshold for detection and sort the ranked list of detected candidates from the entire test set by their saliencies. To plot a new point on the ROC curve, we move down the ranked list by one and look at the portion of the ranked list from its top to the current position, which is equivalent to gradually lowering the global decision threshold. The entire sets of ROC curves computed by this scheme as shown in Fig. 1 are highly densely packed and include every operating point.

Multi-scale saliency approach gives best overall detection performance on both English and Arabic datasets. Using document context, our multi-scale signature detector achieves 92.8% and 86.6% detection rates for the Tobacco-800 and Maryland Arabic datasets, at 0.3 false-positives per image (FPPI). Encouragingly, the advantage of multi-scale approach becomes more obvious on a more diverse dataset, like Tobacco-800. Exploring global context is more effective on machine printed documents as geometric relationships among text lines are more uniform.



(a)

(b)

Figure 1: ROC curves for (a) Tobacco-800 dataset and (b) Maryland Arabic dataset.

Second, we test how discriminative is our proposed saliency measure in capturing the global cursive pattern embedded in signatures. The handwritten Maryland Arabic dataset serves better for this purpose, because variations among local features including size, is not discriminative, as evident from the poor performance of Fisher classifier. Figs. 2 and 3 show samples of detected signatures from Tobacco-800 and Maryland Arabic datasets, together with their saliency maps. We show the top three most salient parts in red, green, and blue, respectively. In our experiment, a cursive structure is normally more than an order of magnitude more salient than printed text of the same dimensions.

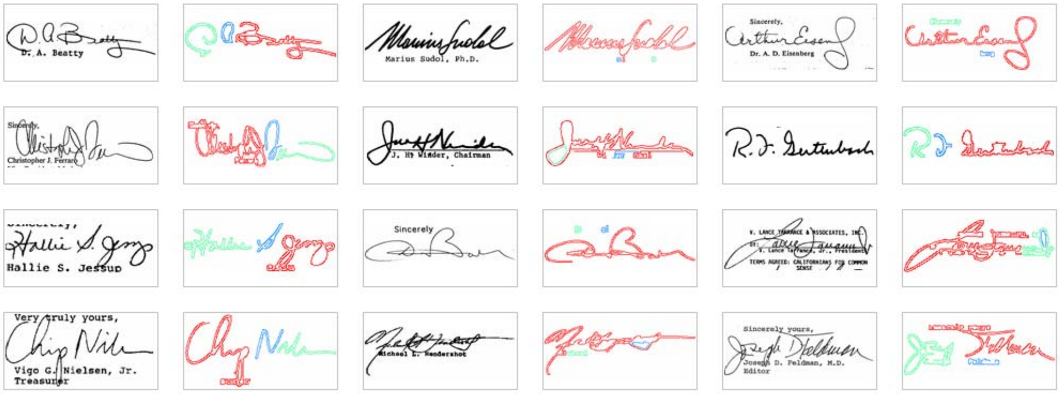


Figure 2: Examples of detected signatures from the Tobacco-800 dataset, together with their saliency maps. The top three most salient parts are shown in red, green, and blue, respectively.

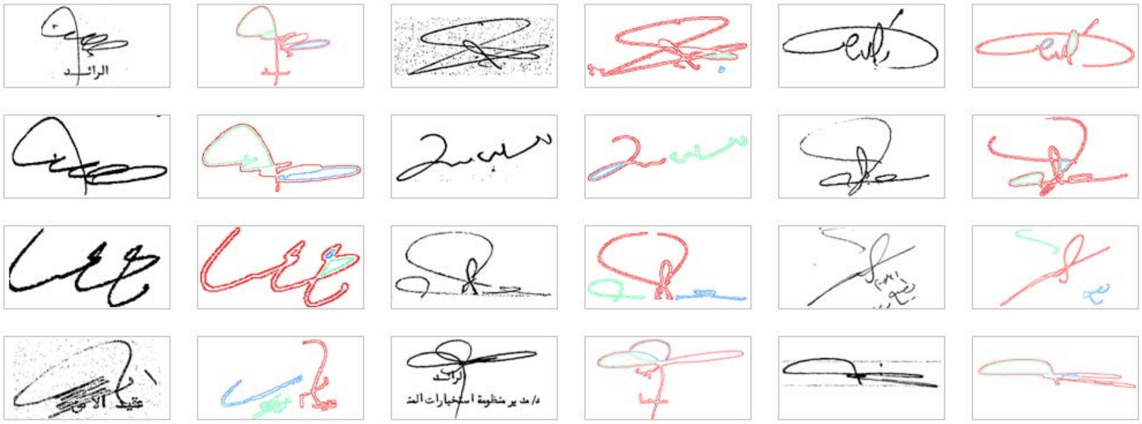
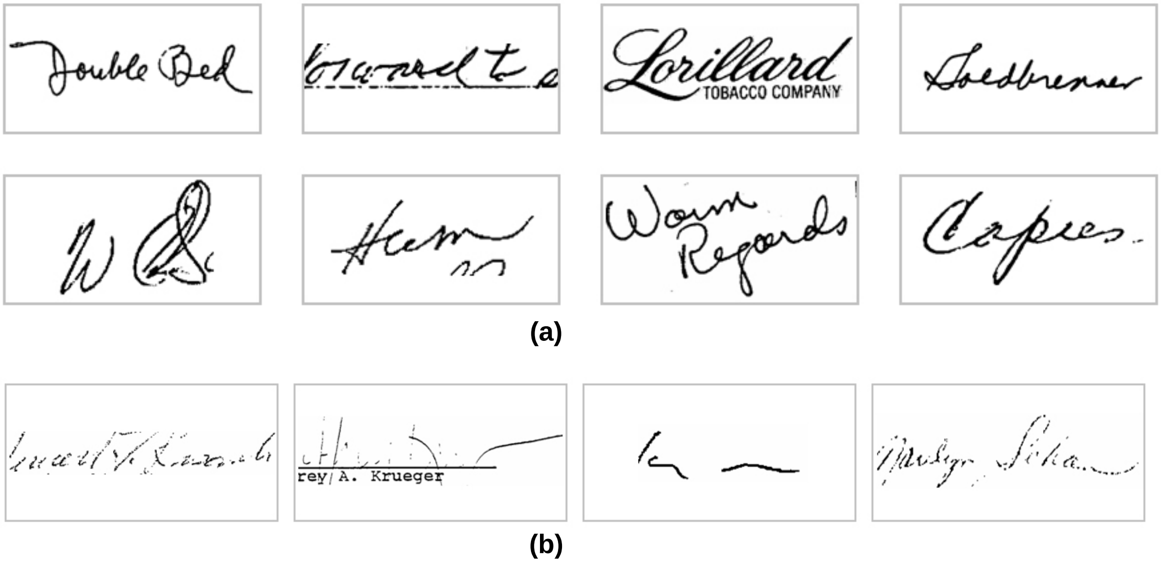


Figure 3: Examples of detected signatures from the Maryland Arabic dataset, together with their saliency maps. The top three most salient parts are shown in red, green, and blue, respectively.

However, we did find a few instances of printed text among false alarms that obtain saliencies comparable to signatures because of their highly cursive fonts, as shown in Fig. 4(a). A limitation of our proposed method is that the detected and segmented signature may contain a few touching printed characters when signatures overlap with very strong background. Nevertheless, the quality of segmented output by structural saliency is considerably better.



(a)

(b)

Figure 4: Examples of (a) false alarms and (b) missed signatures from the Tobacco-800 dataset.

For better interpretation of the overall detection performance, we summarize key evaluation statistics. On Tobacco-800, 848 signatures out of the 900 labeled signatures are correctly detected, by the multi-scale saliency approach using document context in Fig. 1(a). Among correctly detected signatures, 83.3% are complete. Their mean percentage area overlap with the groundtruth is 86.8% with a standard deviation of 11.5%. As shown in Figs. 2 and 3, the quality of detected signatures is comparable to manually cropped versions. This demonstrates that using connected components give extracted signatures of impressive quality, and it does not necessarily limit the detection probability when used in a multi-scale approach. In fact, these figures are close to the machine printed text word segmentation performance level from leading commercial OCR product on Tobacco-800 documents. The results on the Maryland Arabic dataset are also very encouraging as the collection consists mainly of unconstrained handwriting in complex layouts and backgrounds.

4.2 Logo Detection

4.2.1 Datasets

To facilitate a realistic evaluation on logo detection and extraction, we used a large document collection—the Tobacco-800 dataset. Groundtruth of the entire collection was manually created using our developed Java groundtruth editor, and each logo was labeled by a tight rectangular bounding box. We used 50 documents with logos from Tobacco-800 as the training set, and used the rest for testing.

4.2.2 Evaluation Methodology

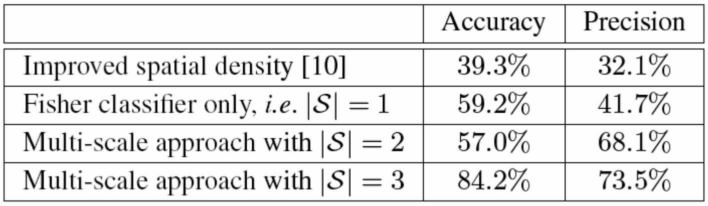
In the following evaluation, we consider a logo correctly detected if and only if the detected region contains more than 75% pixels of a groundtruthed logo and the area of the detected region is less than 125% of the area of that groundtruthed logo. These lower and upper bounds ensure that a detected logo must contain less than 25% missing pixels and less than 25% outliers.

We use accuracy and precision as metrics to evaluate overall detection performance.

4.2.3 Results and Discussion

In this section, we quantitatively evaluate the performance of three different logo detectors: the spatial density approach [4], the Fisher classifier only (i.e. |S| = 1), and the multi-scale detection approach using a varying number of classifiers in cascade. For evaluation purpose, we implemented an improved version of Pham's method, which computes the spatial density using the mountain function for each connected component formed at the initial coarse scale level σ_n. We select σ_n as a linear function within the range of [8, 16] for varying image resolutions. Each scale level deviates from its neighboring scale by a factor of 2. We assume that each document image contains at most one logo, and we run the logo detectors on the top one third of the document. If more than one logo is detected, the one with its projection furthest away from the decision threshold is selected.

Table 2: Summary of logo detection performance.

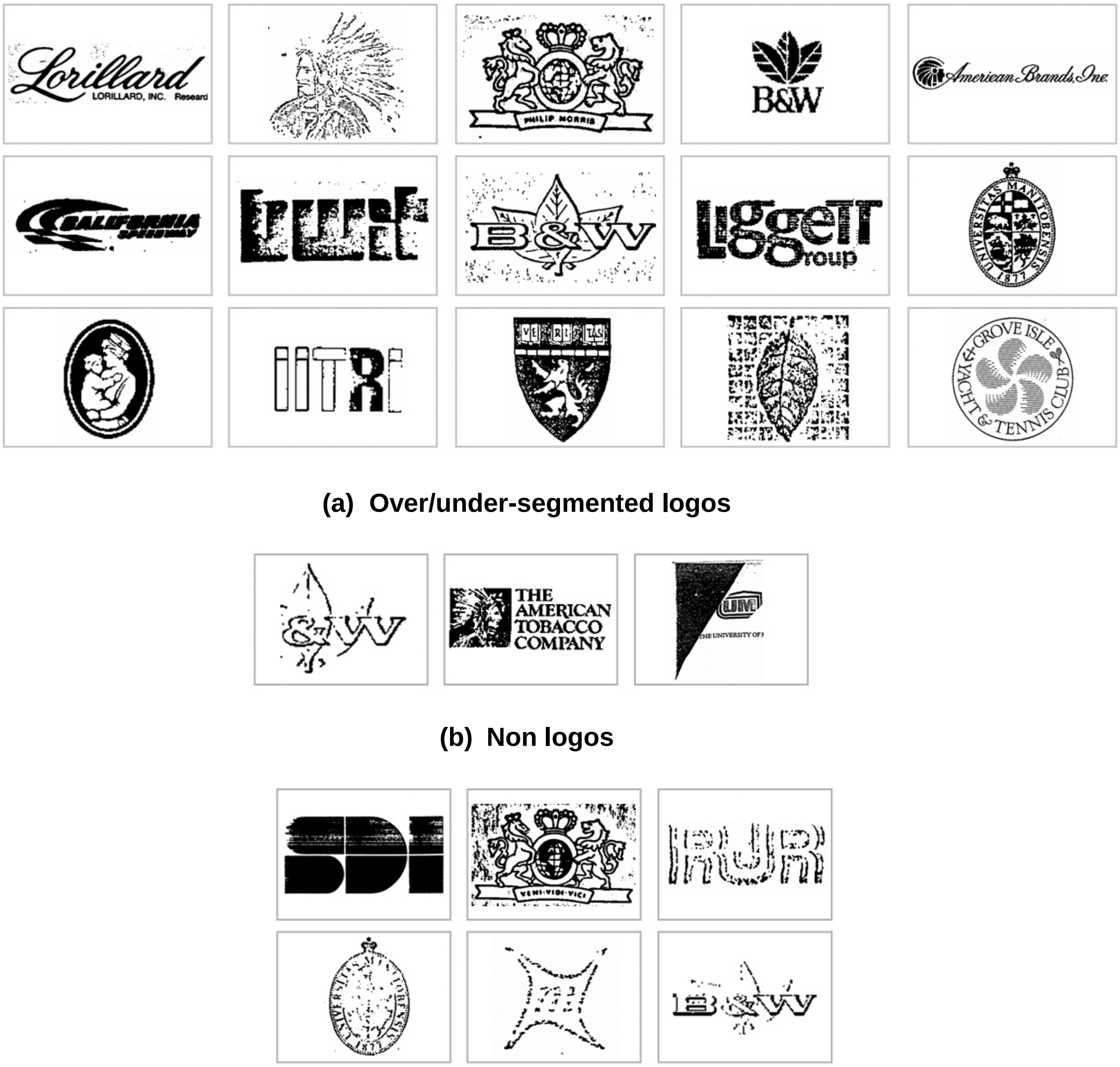


	Accuracy	Precision
Improved spatial density [10]	39.3%	32.1%
Fisher classifier only, *i.e.* $\lvert\mathcal{S}\rvert = 1$	59.2%	41.7%
Multi-scale approach with $\lvert\mathcal{S}\rvert = 2$	57.0%	68.1%
Multi-scale approach with $\lvert\mathcal{S}\rvert = 3$	84.2%	73.5%

Both the Fisher classifier and our multi-scale detection approach perform much better than the improved spatial density method. This is expected because Pham's approach employs only a single feature, and thus is inadequate for real datasets with large variations. Our multi-scale approach gives the best result with |S| = 3, achieving an 84.2% accuracy and 73.5% precision in logo detection. The boosting strategy, through a cascade of classifiers across image scales, is shown to be very effective. Compared to the Fisher classifier, we observe considerable improvement in precision when |S| = 2, as the number of false alarms drops 32.6%. The number of false alarms is further reduced by 44.8% when |S| = 3, which also leads to an increase in accuracy, since we select at most one logo in a document. These results highlight the importance of using a wide range of image scales to effectively tackle large variations in detected objects.

The intuition behind our multi-scale detection approach is in line with the results summarized in Table 2. Linear discriminant analysis provides a low-dimensional subspace that is efficient for separating data. However, if no reasonable spread in data can be assumed (e.g. non-Gaussian), knowing this optimal projection direction is often not sufficient for classifying novel patterns that have large and complex variations with both high accuracy and precision. Combining the Fisher classifier with a cascade of classifiers at multiple scales is an effective approach to pruning the likely data points along the optimal projection direction. This provides an elegant solution to improving both accuracy and precision as the false alarms are dismissed.

The accurate localization of extracted logos using the multi-scale detection approach is evident. On average, a correctly detected logo contains 99.7% of pixels in the logo groundtruth, with a standard deviation of 1.6%. In other words, the quality of these automatically detected and extracted logos is almost identical to that of manually cropped logos. This performance underlines the feasibility of a fully automated logo detection and recognition system.



(a) Over/under-segmented logos

(b) Non logos

(c) Missed logos

Figure 5: Examples of incorrectly detected and missed logos. Both (a) and (b) are considered false alarms in our evaluation.

The multi-scale logo detection algorithm is tractable and highly parallelizable. If we let the total number of pixels in the original image be N, forming connected components requires *O(N)* time. We implemented Gaussian smoothing as two rounds of 1-D convolution using separable kernels. The total complexity in logo detection and extraction algorithm across a total of k scale levels is therefore *O(N)*, with k as a small constant. The average processing time of our serial C++ implementation for a business document scanned at 300 DPI is 0.68 second at $|S| = 3$.

4.3 Stamp Detection

In all tests, we only provided the following a priori knowledge as input to the system, which account for stamp characteristics that can be either reasonably assumed or practically obtainable from limited stamp samples: (a) rough estimates of the minimum and maximum areas of the retrieved stamp, (b) an eccentricity bound of 0.94 (i.e. $a \leq 3b$), which represents the range of eccentricities for normal stamp patterns.

Table 1: Summary of the two test databases of real degraded binary images.

Test Databases	Total Images	Images with The Retrieved Stamp	Image Quality	Stamp Quality	Eccentricity of Retrieved Stamp
Database 1	436	92	Mediate – Poor	Mediate – Poor	Close to 0
Database 2	193	68	Good – Poor	Mediate – Poor	0.71

On each image, top stamp candidates in the three-dimensional parameter space (x_o, y_o, area) are ranked by their scores, which are calculated from their weighted sum of accumulated votes. Strong stamp patterns typically correspond to peaks in the parameter space that are an order of magnitude larger than any other closely ranked top candidates. Once a stamp candidate emerges, we obtain its confidence value by taking the ratio of its scores with the sum of scores from itself and the immediate next top 10 non-stamp candidates. This measure ensures that multiple stamp instances can detected within the same document.

Retrieved stamps from all images are ranked by their confidence values and the detected regions are saved as sub-images for verification. When calculating the mean average precision, we declare the retrieved stamp relevant if the detected region is correct. Figure 6 shows the overall recall-precision trend on the testing data by unconstraint search on entire documents using a priori information described above.

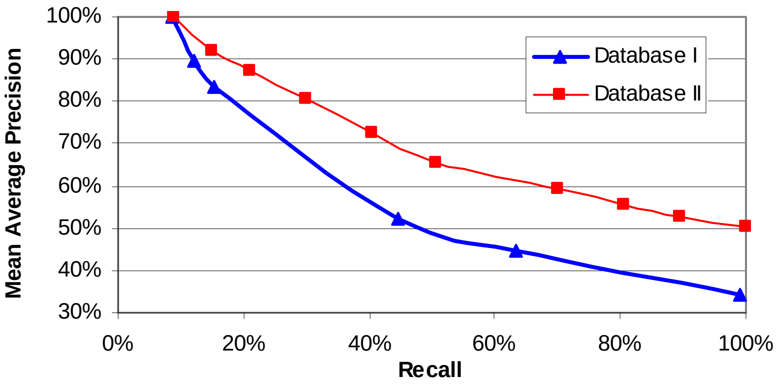


Figure 6: Evaluation of stamp retrievals on groundtruthed document databases.

References

[1] The IIT complex document image processing (CDIP) test collection, http://ir.iit.edu/projects/CDIP.html, 2006.

[2] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. SIGIR, pp. 665–666, 2006.

[3] The legacy tobacco document library (LTDL) at UCSF, http://legacy.library.ucsf.edu/, 2006.

[4] T. Pham. Unconstrained logo detection in document images. *Pattern Recognition*, 36(12), pp. 3023–3025, 2003.

5 Bibliography

Bibliography

Belaïd, A. and L. Pierron "A generic approach for OCR performance evaluation." Proc. SPIE Conference on Document Recognition and Retrieval IX: 203-215.

Blue, J. L., G. T. Candela, et al. (1998). "Evaluation of Pattern Classifiers for Fingerprint and OCR Applications."

Junker, M. and R. Hoch (1998). "An experimental evaluation of OCR text representations for learning document classifiers." International Journal on Document Analysis and Recognition **1**(2): 116-122.

Junker, M., R. Hoch, et al. (1999). "On the evaluation of document analysis components by recall, precision, and accuracy." International Conference on Document Analysis and Recognition: 713-716.

Kanai, J., S. V. Rice, et al. (1995). "Automated evaluation of OCR zoning." IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(1): 86-90.

Kanungo, T., G. A. Marton, et al. "OmniPage vs. Sakhr: Paired Model Evaluation of Two Arabic OCR Products." Proceedings of SPIE Conference on Document Recognition **3651**: 109-120.

Liang, J. (1999). Document Structure Analysis and Performance Evaluation, University of Washington.

Liang, J., I. T. Phillips, et al. "Performance evaluation of document layout analysis algorithms on the UW data set." Proceedings of SPIE Conference on Document Recognition **3027**: 149-160.

Mao, S. and T. Kanungo (2001). "Empirical performance evaluation methodology and its application to page segmentation algorithms." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3): 242-256.

Patton, M. Q. and M. Q. Patton (1987). How to Use Qualitative Methods in Evaluation, Sage Publications Inc.

Trier, O. D. and T. Taxt (1995). "Evaluation of binarization methods for document images." Pattern Analysis and Machine Intelligence, IEEE Transactions on **17**(3): 312-315.

Wang, Y., R. Haralick, et al. (2001). "Zone content classification and its performance evaluation." Proc. ICDAR: 540–544.

6 Dataset Distribution

Genre	
Forms, Drawing, Tables et at.	
Forms	644
Drawing	42
Tables	100
Chemistry formulae	25
Math equations	165
Figures	40
Total	1016
Business documents and Memo letters	
Business documents clean	52
Business documents degraded	2700
Business documents with annotations	160
Memo letters (English + Multilingual)	978
Total	3890
Journal and Conference Papers, Articles	
English	2785
German	359
Japanese	478
Total	3622
Newsletters and Flyers	
Google images	1417
Arabic Newswire + Broadcast News	338
Total	1755
Structured Documents	
Phonebook	229
Dictionaries (Chinese English, English Chinese)	1148
Yellowpage	84
Total	1461
Handwritten	
Arabic	60
Chinese	146
Cyrillic	410
Japanese	47
Korean	80
Thai	319
Hindi	281
Total	1343

Page Classification Datasets (Google Image)	
Document	757
Image with Text	2443
Non-Document	3420
Total	6620
Total in all genre categories	19707