BOBCAT UMD Progress Report
February 14, 2008

Task: Metrics
Researchers: Wontaek Seo, David Doermann
Status:

Algorithms and a frame work has been designed and implemented for document zone classification evaluation. The program takes to DocLib XML files and compares them on a zone by zone basis, assuming zoneids match, and produces

- A file by file evaluation, showing the zones which are correct and incorrect (See Appendix A)
- A summery of accuracy by zone type (See Appendix A)
- A confusion matrix (See Appendix A)
- A Visual output in XML format showing the correct and incorrect images overlaid on the (See Appendix B)

Planned Work:
Challenges and Issues:

None at this time

Task: Survey of Available data and Metrics
Researchers: Kamal Tayal
Status:

A www site has been designed and is operational for the collection of information about datasets, tools and metrics for evaluation. The site will collect contact information, particulars of the datasets and metrics and information about availability and cost.

Planned Work:

February 2008 – testing and finishing implementation
March 2008 – open data collection and targeted advertising

Challenges and Issues:

The timeline looks like the delivery will slip a little from the previously specified February 15th.

Task: Ground Truth Data
Researchers: David Doermann
Status:

A specification has been designed and a preliminary version will be available in about 1 month.

Planned Work:

Ongoing

Challenges and Issues:

None at this time

Task: GEDI Enhancements
Researchers: Elena Zotkina

Status:

The GEDI tool has been modified to provide reading order capabilities. It has also added some automated “box shrinking” capabilities that will be used for zone level ground truthing

Planned Work:

Testing and Delivery of beta version by the end of Feb 2008.

Appendix A: Evaluation Output

```
Zone classification Evaluation Result
Generated on Feb 14 00:20:42 2008

========================
Result of Individual File
========================

O : Matched, X : Mis-matched

A001BIN.TIF
-----------
(O) ZoneID : 000, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 001, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 002, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 003, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 004, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 005, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 006, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 007, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 008, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 009, Ground-Truth : text_sm, Output : text_sm
[OVERALL] 10/10, 100.00%

A002BIN.TIF
-----------
(O) ZoneID : 000, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 001, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 002, Ground-Truth : text_sm, Output : text_sm
(O) ZoneID : 003, Ground-Truth : table, Output : table
[OVERALL] 4/4, 100.00%

<<SKIPPED SOME CONTENT>>

==================
Summary of Results
==================

- Total Number of Sample :  24531
- Overall Accuracy : 94.73%
- Average of Each Class Accuracy : 66.12%

01. Information on Classes
==========================
```

Label	Name of Class	Number of Sample	Accuracy
00	text_sm	21572	95.89%
01	table	204	49.02%
02	math	837	91.40%
03	text_lg	127	69.29%
04	halftone	384	96.61%
05	drawing	774	89.53%
06	chm_drawing	168	65.48%
07	logo	16	6.25%
08	ruling	435	97.70%
09	map	14	0.00%

02. Confusion Matrix

Out\GT	00	01	02	03	04	05	06	07	08	09
00	20686(95.9%)*	10(4.9%)	48(5.7%)	16(12.6%)	0(0.0%)	10(1.3%)	6(3.6%)	2(12.5%)	0(0.0%)	2(14.3%)
01	28(0.1%)	100(49.0%)*	0(0.0%)	0(0.0%)	2(0.5%)	8(1.0%)	0(0.0%)	0(0.0%)	1(0.2%)	0(0.0%)
02	411(1.9%)	11(5.4%)	765(91.4%)*	17(13.4%)	0(0.0%)	18(2.3%)	36(21.4%)	2(12.5%)	1(0.2%)	0(0.0%)
03	15(0.1%)	0(0.0%)	0(0.0%)	88(69.3%)*	1(0.3%)	11(1.4%)	0(0.0%)	5(31.3%)	0(0.0%)	0(0.0%)
04	403(1.9%)	2(1.0%)	0(0.0%)	0(0.0%)	371(96.6%)*	27(3.5%)	0(0.0%)	5(31.3%)	8(1.8%)	3(21.4%)
05	21(0.1%)	81(39.7%)	24(2.9%)	3(2.4%)	10(2.6%)	693(89.5%)*	16(9.5%)	0(0.0%)	0(0.0%)	9(64.3%)
06	3(0.0%)	0(0.0%)	0(0.0%)	3(2.4%)	0(0.0%)	7(0.9%)	110(65.5%)*	1(6.3%)	0(0.0%)	0(0.0%)
07	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	1(6.3%)*	0(0.0%)	0(0.0%)
08	5(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	425(97.7%)*	0(0.0%)
09	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)*

03. Precision and Recall

Class\Eval	precision	recall	detected	correct	total
00	99.55%	95.89%	20780	20686	21572
01	71.94%	49.02%	139	100	204
02	60.67%	91.40%	1261	765	837
03	73.33%	69.29%	120	88	127
04	45.30%	96.61%	819	371	384
05	80.86%	89.53%	857	693	774
06	88.71%	65.48%	124	110	168
07	100.00%	6.25%	1	1	16
08	98.84%	97.70%	430	425	435
09	0.00%	0.00%	0	0	1

Appendix B: Snapshot of results visualization

Note Small Text was incorrectly recognized and classified as a table (Shown in Red)