CLEAT Document Image Database Overview

1. Executive Summary

A large, heterogeneous collection of documents is a fundamental ingredient for research on document image analysis and recognition and the development of information retrieval systems. This document provides an overview of the CLEAT document image database, produced by the Laboratory for Language and Media Processing (LAMP) at the University of Maryland, which contains:

- Multi-lingual binary and grayscale images scanned directly from University of Maryland handwriting collections
- Binary and grayscale images scanned directly from unstructured multi-lingual document image sources, including business correspondence, technical journals and handwritten notes
- Binary and grayscale images scanned directly from structured multi-lingual document image sources, including dictionaries, phonebooks and yellow pages
- Vast amount of images automatically cropped from on-line sources
- Binary images scanned from 1st and other generation photocopies of real forms
- Document images assembled from existing document databases
- All document images tagged with page level attributes for each page
- Software for viewing document images, visualize and edit ground truth information
- Document segmentation ground truth generated on select subset of document images, with each segment zoned and tagged
- Signature and logo detection ground truth manually created on select subset of document images, with each ground truth region zoned and tagged
- Signature and logo detection results generated by algorithms developed by University of Maryland on select subset of document images, with each detected region zoned and tagged

To accommodate diverse needs in document image analysis and understanding research, the CLEAT document image database includes images with a rich blend of imaging resolutions and degradations.

This volume of the database contains 16,854 images, utilizing a total of more than 6.0 gigabytes of storage. Table 1 provides a detailed description of the CLEAT database in terms of genre type and language distributions.

Table 1: Genre types and distribution of the CLEAT document image database.

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
Yellowpages	84
Total	1461
Handwritten	
Arabic	60
Chinese	146
Cyrillic	410
Japanese	47
Korean	80
Thai	319
Hindi	281
Total	1343
Page Classification Datasets	
Document	797
Image with Text	1695
Non-Document	1275
Total	3767
Total in all genre categories	16854

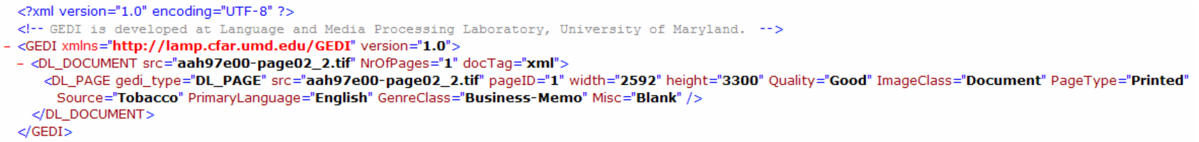
2. Page-level Information

Page-level attributes provide essential information associated with each document image. They are manually created for the entire CLEAT database, and can be further edited using the GEDI software provided.

Table 2: Attribute and value sets used in the ground truth of CLEAT database.

Attributes	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7	Value 8
ImageClass	Document	ImageWithText	Non-Document					
PageType	Printed	Handwritten	Mixed					
GenreClass	Business-Memo Letter Form Other	Article-Table Figure Other	Graphics-Drawing	Structured-Phonebook Yellowpage Dictionary Other	Newsletters	Other-BroadcastNews Newswire	Unknown	
PrimaryLanguage	English	German	Chinese	Japanese	Korean	Thai	Cyrillic	Arabic
Source	Tobacco	Web	ScannedMedia	Other				
Quality	Good	Poor						
Misc								

Table 2 lists the set of common attributes and their associated values used in the creation of CLEAT page-level ground truth. Figure 2 shows an example XML ground truth file displayed using browser.



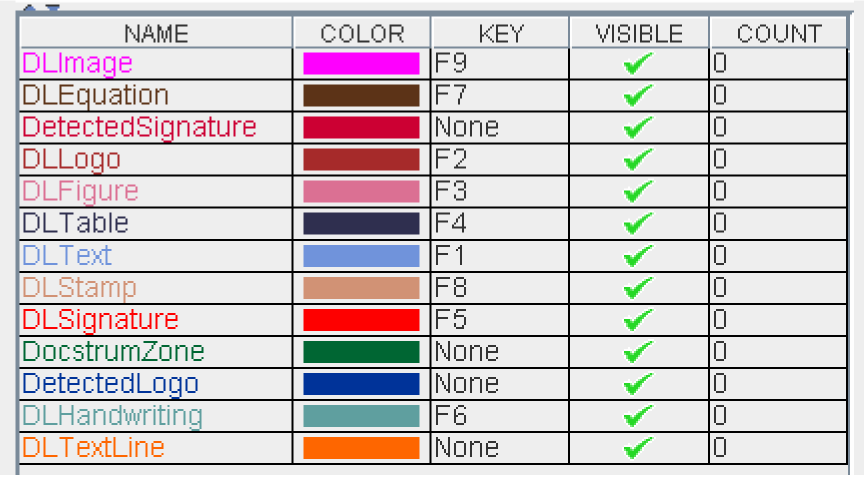
```
  <?xml version="1.0" encoding="UTF-8" ?>
  <!-- GEDI is developed at Language and Media Processing Laboratory, University of Maryland.  -->
- <GEDI xmlns="http://lamp.cfar.umd.edu/GEDI" version="1.0">
  - <DL_DOCUMENT src="aah97e00-page02_2.tif" NrOfPages="1" docTag="xml">
      <DL_PAGE gedi_type="DL_PAGE" src="aah97e00-page02_2.tif" pageID="1" width="2592" height="3300" Quality="Good" ImageClass="Document" PageType="Printed"
        Source="Tobacco" PrimaryLanguage="English" GenreClass="Business-Memo" Misc="Blank" />
    </DL_DOCUMENT>
  </GEDI>
```

Figure 1: Visualization of page-level attributes in XML format.

3. Zone-level Information

Zone-level attributes provide essential information related to a specific region on a document page. The list of zone-level attributes can be defined in an extensible fashion by an application end user. This tight integration with the end application enables training and evaluation of various document image analysis and recognition algorithms on the CLEAT database.

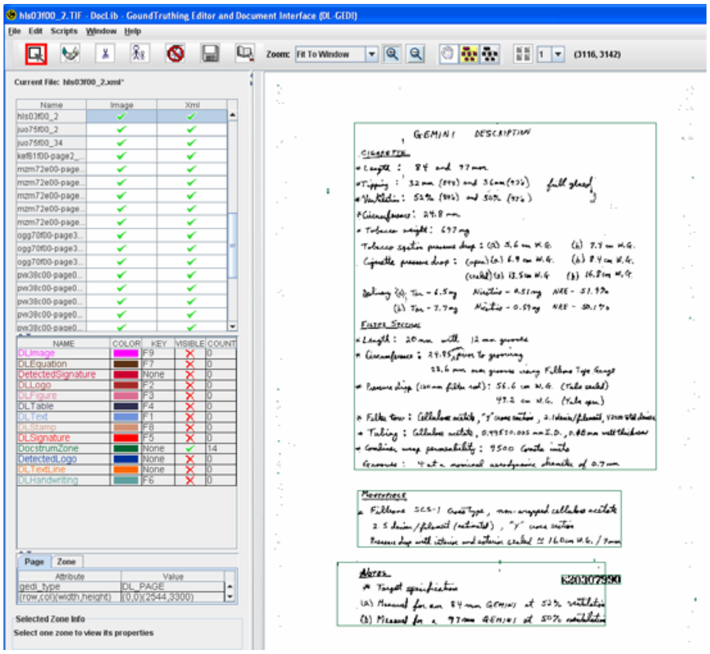
Figure 2 shows a list of zone types, in which the "*DL*" prefix indicates the zone types in the ground truth data.



NAME	COLOR	KEY	VISIBLE	COUNT
DLImage	🟪	F9	✓	0
DLEquation	🟫	F7	✓	0
DetectedSignature	🟥	None	✓	0
DLLogo	🟥	F2	✓	0
DLFigure	🟥	F3	✓	0
DLTable	🟦	F4	✓	0
DLText	🟦	F1	✓	0
DLStamp	🟧	F8	✓	0
DLSignature	🟥	F5	✓	0
DocstrumZone	🟩	None	✓	0
DetectedLogo	🟦	None	✓	0
DLHandwriting	🟩	F6	✓	0
DLTextLine	🟧	None	✓	0

Figure 2: Visualization of the list of extensible zone types, which include pre-defined zone types in the ground truth and those later defined by the user themselves.

Figure 3 shows specific zone types displayed using GEDI software.



hls03f00_2.TIF - DocLib - GoundTruthing Editor and Document Interface (DL-GEDI)

Current File: hls03f00_2.xml*

Name	Image	Xml
hls03f00_2	✓	✓
juo75f00_2	✓	✓
juo75f00_34	✓	✓
kef61f00-page2_...	✓	✓
mzm72e00-page...	✓	✓
mzm72e00-page...	✓	✓
mzm72e00-page...	✓	✓
mzm72e00-page...	✓	✓
mzm72e00-page...	✓	✓
ogg70f00-page3...	✓	✓
ogg70f00-page3...	✓	✓
ogg70f00-page3...	✓	✓
pvx38c00-page0...	✓	✓
pvx38c00-page0...	✓	✓
pvx38c00-page0...	✓	✓
pvx38c00-page0...	✓	✓
pvx38c00-page0...	✓	✓

NAME	COLOR	KEY	VISIBLE	COUNT
DLImage	🟪	F9	✗	0
DLEquation	🟫	F7	✗	0
DetectedSignature	🟥	None	✗	0
DLLogo	🟥	F2	✗	0
DLFigure	🟥	F3	✗	0
DLTable	🟦	F4	✗	0
DLText	🟦	F1	✗	0
DLStamp	🟧	F8	✗	0
DLSignature	🟥	F5	✗	0
DocstrumZone	🟩	None	✓	14
DetectedLogo	🟦	None	✗	0
DLTextLine	🟧	None	✗	0
DLHandwriting	🟩	F6	✗	0

Page | Zone

Attribute	Value
gedi_type	DL_PAGE
(row,col)(width,height)	(0,0)(2544,3300)

Selected Zone Info

Select one zone to view its properties

GEMINI DESCRIPTION

CIGARETTE

* Length : 84 and 77 mm
* Tipping : 32 mm (84) and 36 mm (77) full glued
* Ventilation : 52% (84) and 50% (77)
* Circumference : 24.8 mm
* Tobacco weight : 677 mg

Tobacco section pressure drop : (a) 5.6 cm W.G. (b) 7.7 cm W.G.
Cigarette pressure drop : (open) (a) 6.9 cm W.G. (b) 8.4 cm W.G.
(sealed) (a) 13.5 cm W.G (b) 16.8 cm W.G
Delivery (a) Tar - 6.5 mg Nicotine - 0.51 mg NRE - 51.9%
(b) Tar - 7.7 mg Nicotine - 0.59 mg NRE - 50.1%

FILTER SECTION

* Length : 20 mm with 12 mm grooves
* Circumference : 24.95 prior to grooving
23.6 mm over grooves using Filtrona Type Gauge
* Pressure drop (120 mm filter rod) : 56.6 cm W.G. (Tube sealed)
47.2 cm W.G. (Tube open)
* Filter tow : Cellulose acetate, "Y" cross section, 2.1 denier/filament, [illegible] total denier
* Tubing : Cellulose acetate, 0.995±0.005 mm I.D., 0.18 mm wall thickness
* Combiner wrap permeability : 9500 Coresta units

Grooves : 4 at a nominal aerodynamic diameter of 0.7 mm

MOUTHPIECE

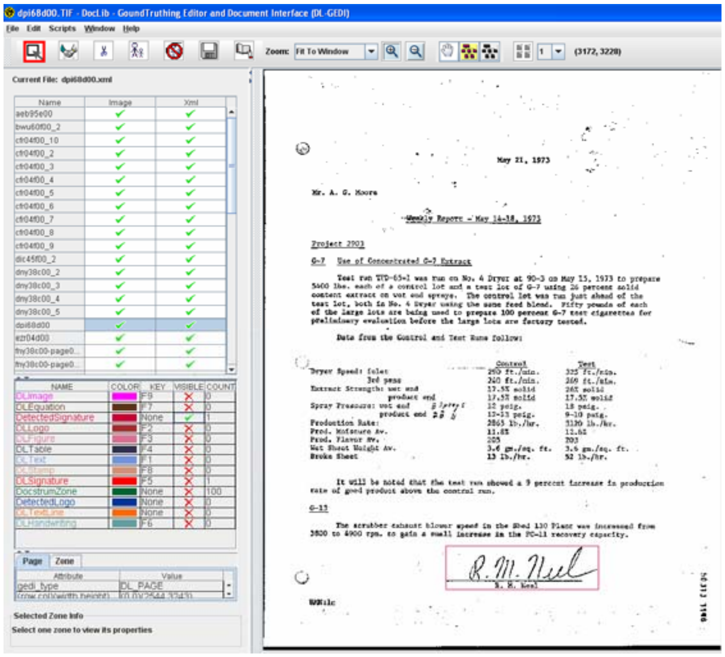
* Filtrona SCS-1 Cross Type, non-wrapped cellulose acetate
2.5 denier/filament (estimated), "Y" cross section
Pressure drop with interior and exterior sealed ≈ 16.0 cm W.G. / 7 mm

NOTES

620307990

* Target specification
(a) Measured for an 84 mm GEMINI at 52% ventilation
(b) Measured for a 77 mm GEMINI at 50% ventilation

(a)


dpi68d00.TIF - DocLib - GoundTruthing Editor and Document Interface (DL-GEDI)
File Edit Scripts Window Help
Zoom: Fit To Window
1
(3172, 3228)
Page
Zone

Current File: dpi68d00.xml

Name	Image	Xml
aeb95e00	✓	✓
bwu60f00_2	✓	✓
cfr04f00_10	✓	✓
cfr04f00_2	✓	✓
cfr04f00_3	✓	✓
cfr04f00_4	✓	✓
cfr04f00_5	✓	✓
cfr04f00_6	✓	✓
cfr04f00_7	✓	✓
cfr04f00_8	✓	✓
cfr04f00_9	✓	✓
dic45f00_2	✓	✓
dny38c00_2	✓	✓
dny38c00_3	✓	✓
dny38c00_4	✓	✓
dny38c00_5	✓	✓
dpi68d00	✓	✓
ezr04d00	✓	✓
fny38c00-page0...	✓	✓
fny38c00-page0...	✓	✓

NAME	COLOR	KEY	VISIBLE	COUNT
DLImage	■	F9	✗	0
DLEquation	■	F7	✗	0
DetectedSignature	■	None	✓	1
DLLogo	■	F2	✗	0
DLFigure	■	F3	✗	0
DLTable	■	F4	✗	0
DLText	■	F1	✗	0
DLStamp	■	F8	✗	0
DLSignature	■	F5	✗	1
DocstrumZone	■	None	✗	100
DetectedLogo	■	None	✗	0
DLTextLine	■	None	✗	0
DLHandwriting	■	F6	✗	0

Attribute	Value
gedi_type	DL_PAGE
(row,col)(width,height)	(0,0)(2544,3243)

Selected Zone Info

Select one zone to view its properties

May 21, 1973

Mr. A. G. Moore

Weekly Report – May 14-18, 1973

Project 2903

G-7 Use of Concentrated G-7 Extract

Test run TTD-65-1 was run on No. 4 Dryer at 90-3 on May 15, 1973 to prepare 5400 lbs. each of a control lot and a test lot of G-7 using 26 percent solid content extract on wet end sprays. The control lot was run just ahead of the test lot, both in No. 4 Dryer using the same feed blend. Fifty pounds of each of the large lots are being used to prepare 100 percent G-7 test cigarettes for preliminary evaluation before the large lots are factory tested.

Data from the Control and Test Runs follow:

	Control	Test
Dryer Speed: Inlet	250 ft./min.	325 ft./min.
3rd pass	240 ft./min.	309 ft./min.
Extract Strength: wet end	17.5% solid	26% solid
product end	17.5% solid	17.5% solid
Spray Pressure: wet end 8 sprays	12 psig.	18 psig.
product end 88 8	12-13 psig.	9-10 psig.
Production Rate:	2865 lb./hr.	3120 lb./hr.
Prod. Moisture Av.	11.8%	12.6%
Prod. Flavor Av.	205	203
Wet Sheet Weight Av.	3.6 gm./sq. ft.	3.6 gm./sq. ft.
Broke Sheet	13 lb./hr.	52 lb./hr.

It will be noted that the test run showed a 9 percent increase in production rate of good product above the control run.

G-13

The scrubber exhaust blower speed in the Shed 130 Plant was increased from 3800 to 4900 rpm. to gain a small increase in the PC-11 recovery capacity.

R. M. Neel
R. M. Neel

RMN:lc

50313 1146

(b)

Figure 3: Display of zone with select types using GEDI software (a) Page segments. (b) Detected signature regions.

4. Contact Information

For ordering information contact:

```
Laboratory for Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, Maryland 20742

Attention: Dr. David Doermann

Phone: 301-405-1767
FAX: 443-638-0236
E-mail: doermann@umiacs.umd.edu
http://lamp.cfar.umd.edu
```