BOBCAT UMD Progress Report
May 30, 2008

Task: Metrics
Researchers: Wontaek Seo, David Doermann

Completed Work:

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)

We have implemented a doclib module for zone segmentation evaluation found in the following ICDAR article.

- Antonacopoulos, A.; Gatos, B.; Karatzas, D., "ICDAR 2003 page segmentation competition," *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on* , vol., no., pp. 688-692, 3-6 Aug. 2003

Software Overview:

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:

Student has left the University and we are currently seeking a replacement.

Task: Ground Truth Data
Researchers: David Doermann

Status:

50 sample documents have been annotated at the line level from the ANFAL data and are being used for evaluation.

Planned Work:

Ongoing

Challenges and Issues:

None at this time

Task: GEDI Enhancements
Researchers: Elena Zotkina and David Doermann

Previous Work:

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

Enhancements include

- File Read/Write Warnings
- Reading Order Capabilities
- Network listener added to control GEDI from outside the program
- Linux Support
- Multiple File Type Support

Version 2.1 tested and delivered April 15th.

Gedi most recently has been enhanced to allow line level alignment of ground truth data, easy navigation between regions, linking of reading order to annotation.

Planned Work:

We are in the process of building in performance evaluation visualization.

Appendix A: Functional Description of Performance Evaluation Tool of Zone Segmentation Algorithm

Overview :

PEZS is command line based program which evaluates the performance of zone segmentation algorithsm. This program is written using C++ language on the UINX/LINUX platform and is being ported to Windows platform in the near future. For the performance evaluation of zone segmentation algorithm, this program needs three files as input, an image file of document, a ground truth file which follows the GEDI format and a result file which also follows the GEDI format. In the GEDI format file, Document, Page and Zone are described as XML file format. A zone is defined by three types of box, rectangle, rectangle with orientation and polygon. A rectangle box has 4 attributes to represent the box, two x, y coordinates represented as col and row, width and height. A rectangle box with orientation has one more attribute than a rectangle box which is the degree of orientation of the box. A polygon box only has a set of point to be connected. Every zone has a "gedi_type" attribute that represent the label of zone. In the program, every zone in the result file is matched to every zone in the ground truth file. This program uses a matching score that is calculated using ON pixel counting. ON pixels are considered foreground pixels which are black pixel on the document image. To compare every zone from result file and ground truth file, matching score table is constructed, and the matching score which is bigger than a threshold is considered to compare the label. If one result zone has a high matching score and has same label with a zone from ground truth, this result zone could be defined as MATCH zone. If one result zone has a high matching score and has different label, then this result zone is defined as DETECT zone. If a result zone has no high matching score to any ground truth zone, this zone is defined as FALSEALRAM zone. And if a ground truth zone has no high matching score to any result zone, this ground truth is defined as MISSED zone. This kinds of result is exported to the output file.

Name :

PEZS – Performance Evaluation tool of Zone Segmentation

Synopsis :

```
PEZS    -r { <file>| <dir> } -g { <file> | <dir> } -img { <file> | <dir> }
        [ -o <file> -detail -v <dir> -t NUM ]
```

Options :

-r { <file> | <dir> } : Path to the result file or directory that has multiple result files of the segmentation algorithm.

-g { <file> | <dir> } : Path to the ground truth file or directory

-img { <file> | <dir> } : Path to the document image file or directory

-o <file> : Filename which the evaluation results are going to be saved

-detail : When this option is activated, detailed result of each zone will be added in the result file, otherwise summary part of result will be exported to the result file.

-v <dir> : Path to the directory which the GEDI format result for visualization will be saved

-t NUM : set the threshold by user, otherwise 80% will be used.