Action Plan Background: PRO

Author: Carol Chou
Release Date: May 4, 2007
Last Revision: Oct 10, 2007

Change History:
10/10/2007 PRO file is no longer a subtype of plain text file format. Change MIME media type to "application".

Preface:

PRO (Prime Recognition Format) is a proprietary format developed by Prime Recognition, Inc. It is the native output format generated by PrimeOCR, optical character recognition (OCR) software, for recording OCR results. The format specification for the PRO file format is described in section six of the PrimeOCR Access Guide [1], PrimeOCR Output Formats.

1 General Description

1.1 Format Name: PRO file format.

1.2 Version: Each PRO file contains version information that specifies the version of the PrimeOCR software which outputs the PRO file. The current version of the PrimeOCR software is 4.30.

1.3 MIME media type name: PRO is a custom format generated by PrimeOCR, so its MIME media type is "application".

1.4 MIME subtype: Following the IANA MIME type rule of prefixing 'x' for unregistered mime types, PRO format would use 'x-pro' as its MIME subtype.

1.5 Short Description: PRO contains all characters recognized by PrimeOCR. For each recognized character, it describes the character attributes, size and a bounding box specifying the area enclosing the character.

1.6 Common Extensions: .pro. PrimeOCR uses ".pro" as the default file extension for PRO files; however, it is not a requirement to use ".pro" file extension.

1.7 Color depth: N/A

1.8 Color Space: N/A

1.9 Compression: N/A

1.10 Progressive Display: N/A

1.11 Animation: N/A

1.12 Magic number(s): The first line of a PRO file must contain a **"**Version, Format, Confidence**"** string somewhere in the line. Earlier versions of PrimeOCR, version 3.0 and older, use "PR Version Number, Word Output, Average Confidence" instead. The content in the PRO format specification is outdated and still refers to the text used in older versions of PrimeOCR [9].

1.13 Specification Requirements:
A PRO file is a document describing the PrimeOCR output. It contains all pages processed by PrimeOCR. Each page of the document must begins with five lines of text describing the page settings. The first line describes the PrimeOCR version used and the average confidence of all characters on this page. The second line contains user tag information, followed by the file name of the original image. The fourth line contains the page configuration data such as page quality, DPI, language, image size and resolution. The last line describes the number of zones in this page.

There must be at least one zone defined in each page. A zone is described in two lines where the first line contains zone configuration settings followed by a line specifying the number of lines in the zone. Each recognized line is described using nine lines of text. These text are used to describe the recognized characters with the associated confidence level, attributes, size and a bounding box for each character.

Each page is ended with a '\f' character (end of page).

2 Essential and Distinguishing Characteristics

PRO is a file format for storing OCR results generated from PrimeOCR. It contains characters recognized during OCR processes, plus the OCR configuration settings used. Each recognized character is associated with a bounding box enclosing the character on the original image. Bounding boxes are in units of 1/1200 inch.

PrimeOCR supports a wide range of fonts. It adopts the omnifont technology which can potentially recognize any font with standard character attributes. In addition, PrimeOCR includes multi-lingual support and can recognize most ANSI characters. Some special characters such as "·" (bullet) are mapped and are output in different characters, in this case "*". Characters in Asian fonts are output in Unicode format. Appendix A of the PrimeOCR Access Guide provides a mapping of all recognized characters. The mapping is consistent with the ISO 8859-x standard.

2.1 PRO Technical Metadata

Technical Metadata Element (G = general file metadata, GT = general text metadata, F = format specific medatadata)	**Obligation (R = Required by spec., S= Defined in spec., D = Derived from spec., O = Optional)**
Number of pages	D
Average confidence of all recognized characters in the page [F]	R
Recognized language	R
Page Quality (1-9)	R

DPI (Dots Per Inch)	R
Width of the original image (BMU)	R
Height of the original image (BMU)	R
Number of zones in each page [F]	R
Number of lines in each zone [F]	R

Note: BMU is defined as 1/1200 inch in PRO

3 Usefulness

3.1 Version Duration:

The most recent version of PrimeOCR is 4.30, released on September 2006. It has been seven months since the current version was released.

3.2 History of Prior Versions Duration:

PrimeOCR was originally released in 1995. Since its original debut, Prime Recognition has made several revisions of its PrimeOCR software. Unfortunately, there is no information about how often PRO file format is revised. For PrimeOCR, the following revision history is available on the corporate news of Prime Recognition websites.

Version 2.5 and prior – unknown.
Version 2.6 – released on April 15, 1997 [2]. This version includes supports for Rich Text Format (RTF), automatic image rotation and graphic export features.
Version 2.7 – released Aug 22, 1997 [2]. Included support for long file name and consolidated logging.
Version 3.0 – released on Feb 16, 1998 [2]. Added support for PDF output and improvement on OCR accuracy.
Version 3.6 – Added support for HTML and comma delimited output. Added lexical check for improving the confidence and accuracy of recognized characters.
Version 3.8 – Added support for color images and window 2000.
Version 3.9 – Added support for PDF bookmark creation and optimization on PDF output.
Version 4.0 – released in January 2005 [3]. Added support for Asian languages, thumbnails in PDF output, and Window 2003 server.
Version 4.2 – released in January 2006 [4]. Added support for LZW compressed images and some performance improvement.
Version 4.3 – released in September 2006. Added PDF/A-1b output and automatic document language identification.
Version 4.4 – release in August 2007. Added support for PDF/A-1a and JBIG2 compression of PDF output. Added sanity check to ensure certain unprintable control characters such as DEL (0x7f) to not be included in the PRO output.

3.3 Expected Newer Versions:

None expected at this moment.

3.4 Existence of Publicly Available Complete Specifications:

The format specification for PRO file format is included as part of the technical specification for PrimeOCR software, section six of PrimeOCR Access Guide version 4.30: PRO Output,

available on Prime Recognition's websites. Prime Recognition updates the PrimeOCR Access Guide along with the releases of its PrimeOCR software. Older versions of PrimeOCR technical specification are not available.

3.5 Specifications-controlling Body:
The format specification is solely controlled by Prime Recognition, Inc. Prime Recognition indicates in PrimeOCR Access Guide that it "reserves the right to modify or revise all or part of this document without notice and shall not be responsible for any loss, cost, or damage, including consequential damage caused by reliance on these materials".

3.6 Related Legal Issues:
There is no license requirement for using PRO file format [8]. The technical contacts in Prime Recognition indicated that they are not aware of any licensing requirement for reading or writing PRO files.

3.7 Application and Platform Support:
PRO file format is mostly text based; most of its content can be rendered in any text editor which supports the encoded characters. However, it may contain unprintable obscure control characters such as DEL (hex: 7F) that may not be renderable in text editors. PRO is a file format used exclusively by Prime Recognition Inc. In addition to PrimeOCR, Prime Recognition has developed other software including PrimePost [5] and PrimeVerify [6] to process PRO files. It is also possible that PrimeOCR users implement custom in-house software to process PRO files for additional features such as search hit highlighting. It is unclear how many in-house software programs may exist. Currently, there appear to be no other commercial products using PRO file format.

PrimeOCR, PrimePost and PrimeVerify are only available on Windows platform.

3.8 Limitations:
A PRO file is simply a file describing OCR result. The limitations in PRO format are more bounded by the limitations in the PrimeOCR software. For PrimeOCR, only one language may be recognized per page. Therefore, a PRO file could contain text in many different languages but each page is limited to only one language. Additionally, each page is limited to have a maximum of 1296 zones where each zone can have up to 500 lines with up to 500 characters per line.

3.9 Perceived Popularity:
PRO is an OCR output format appearing to be generated exclusively by PrimeOCR. PrimeOCR also generates other formats such as PDF, ASCII text, HTML, etc. It is unknown how many institutions export OCR results in PRO files.

For OCR software, there are other popular Desktop OCR software such as ABBYY FineReader, ScanSoft Omnipage, Readiris, Presto OCR and OCRopus (open-source). Figure 1 shows the results of using Google to search the names of these OCR software products. Please note that these results are subject to the search terms used. It does not necessary reflect the popularity of these OCR products, but rather a possible indication of relative popularity among these OCR software products.

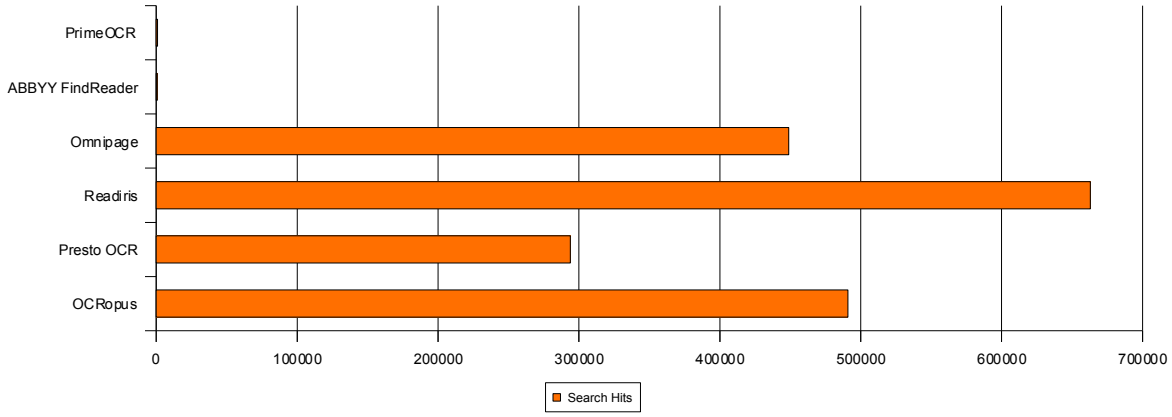

Figure 1: Number of search hits returned by Google for searches performed on April 19, 2007. Each of the Y axes represents the search term used to search the OCR software. There are 884 hits for PrimeOCR and 809 hits for ABBYY FindReader.

Based on the search results from Google, it does not appear that PrimeOCR is as popular as other Desktop OCR software products such as Omnipage or Readiris. PrimeOCR is targeted more toward large-scale OCR processing. It claims to have higher OCR accuracy, 65-80% fewer errors than competing software, by implementing a "voting" system with multiple OCR engines. PrimeOCR is used by some university libraries including the University of Florida and the University of North Texas; government agencies like United State Department of Defense, US National Library of NIH and Department of Commerce; and many corporate offices [7].

4 Related Formats

4.1 Specification Variations:

PRSTAR – This is a custom version of PRO format implemented by Prime Recognition. It is one of the output formats for PrimeOCR. Prime Recognition does not publish the specification for their PRSTAR file format.

5 Summary and Conclusions

PRO format is a custom file format for storing and processing OCR results. It is used exclusively by Prime Recognition, Inc. It is neither a popular nor a standardized file format for OCR output. More popular formats like PDF or HTML are used by most OCR software as their output formats. PrimeOCR supports many other popular formats in addition to PRO. Because PRO contains descriptions of bounding boxes and OCR configuration settings, some institutions using PrimeOCR may choose to archive PRO files as well.

One potential issue that may arise for preserving PRO files is the possibility of an outdated magic number used for format identification. As Prime Recognition may change PRO format without updating PRO format specification, the change in PRO format may not be detected and thus cause failures in PRO format identification. Special care needs to be taken when upgrading to newer versions of PrimeOCR software; making sure the generated PRO files still use the same magic number. To ensure proper format identification for PRO, it is recommended that users notify the FDA when upgrading to a new version of PrimeOCR.

6 References

[1] PrimeOCR Access Guide Version 4.30, Prime Recognition, Inc.
http://www.primerecognition.com/ftp/Manuals/PrimeOCR/PrimeOCR_manual.pdf

[2] News at Prime Recognition, Inc., http://www.primerecognition.com/augprime/news.htm

[3] Prime Recognition Announces PrimeOCR Version 4.0, January 21, 2005
http://press.arrivenet.com/technology/article.php/566668.html

[4] Prime Recognition Announces PrimeOCR Version 4.2, January 20, 2006,
http://openpr.com/news/3527/Prime-Recognition-Announces-PrimeOCR-Version-4-2.html

[5] Other Prime Recognition Products and Services, PrimePost
http://www.primerecognition.com/ftp/Literature/DataSheets/other_products.pdf

[6] PrimeView, PrimeVerify User Guide, Version 4.20,
http://www.primerecognition.com/ftp/Manuals/PrimeProof/PrimeProof_manual.pdf

[7] Prime Recognition Customers, http://www.primerecognition.com/augprime/customer.htm.

[8] Alex Dahl, Email Correspondence, Technical Support, Prime Recognition, Inc., April 27, 2007.

[9] Alex Dahl, Email Correspondence, Technical Support, Prime Recognition, Inc., February 1, 2007.