VIVIDATA IMAGING SOFTWARE FOR UNIX & LINUX

OCR Shop® XTR™ Lite

User's Guide

OCR Shop XTR™ Lite quickly and accurately converts document images into readable text in Linux and Unix environments.

For version 6.0

Copyright Notices

If this Software was acquired under a GSA Schedule, the U.S. Government has agreed to refrain from changing or removing any insignia or lettering from the Software or the accompanying written materials that are provided or from producing copies of manuals or disks (except for backup copies) and:

(e) Title to and ownership of this Software and documentation and any reproductions thereof shall remain with Vividata LLC.
(f) Use of the Software and documentation shall be limited to the facility for which it was acquired except under special contract, and:
(g) If use of the Software is discontinued to the installation specified in the purchase/delivery order and the U.S. Government desires to use it at another location (except under special contract), it may do so giving prior written notice to Vividata LLC, specifying the type of computer and the new site. U.S. Government personnel using this Software, other than under DoD contract or GSA Schedule, are herby on notice that use of this Software is subject to restrictions which are the same as or similar to those specified above.

Request for Comments

In our effort to provide you with the best documentation possible, we welcome any comments and suggestions you may have about our products. Please direct communications to:

Vividata LLC
1300 66th Street
Emeryville, CA 94608
U.S.A.

Phone: (510) 658-6587

www.vividata.com
Send email through our website by clicking Contact Support or Contact Sales.

Table of Contents

Chapter 1: Before You Begin

OCR Shop® XTR™ Lite Overview

OCR Shop® XTR™ Lite quickly and accurately converts scanned images, PDF files, and faxes into ASCII or Unicode text files.

The technology used is Optical Character Recognition (OCR). During OCR, OCR Shop XTR Lite looks for and defines characters in an image to produce text that you can revise without retyping. The underlying OCR engine is based on Vividata's OCR Shop XTR™ /API, utilizing Vividata image processing technology along with ScanSoft® OCR technology. OCR Shop XTR Lite offers a subset of the comprehensive features found in the full OCR Shop XTR, as discussed in Chapter 6.

OCR Shop XTR Lite provides a simple command-line interface to efficiently convert document images into editable text files. Recognizing text in over 50 different languages, OCR Shop XTR Lite offers both ASCII and Unicode output for international support.

For automated applications or batch processing, OCR Shop XTR Lite may be programmatically invoked from applications or scripts. For example, programs written in Java or C/C++, shell scripts, and scripting languages such as Perl, PHP, and Python can all be used to develop applications that use OCR Shop XTR Lite.

Type Conventions

Different kinds of typefaces used throughout this manual indicate text that will appear on the screen or need to be entered by the user.

Type:	Indicates text is:
`courier`	`text generated by the computer`
`courier bold`	**`text typed in by user`**
<brackets>	text to be replaced by user

When asked to enter commands preceded by a pound sign (`#'), the user should be in super-user mode or `root' first. (The command to be entered does not include the pound sign itself.)

System Requirements

OCR Shop XTR Lite is available for a variety of Unix-based workstations. The following platforms are currently supported:

Manufacturer	Operating System / CPU
Sun/Oracle	Solaris SPARC (Solaris™ 2.10+)
Linux: RedHat, Debian, etc.	Linux x86 (Kernel 2.6 and higher)

If your platform is not listed above, you can contact Vividata LLC to see if your platform has been added since this printing of the manual.

Customer Support

You can reach the Vividata LLC technical support staff by:

www.vividata.com – Click 'Contact Support' to send us an email.

Telephone: USA (510) 658-6587
Fax: USA (510) 658-6597

Customer Service is available on regular business days during normal business hours (U.S.A. Pacific time).

Chapter 2: Software Installation

Overview

This section describes how to install the OCR Shop XTR Lite software and license. Consult the release notes for any last-minute information relevant to your particular system. If you need to download the software evaluation, go to www.vividata.com.

Installing OCR Shop XTR Lite

1. Place the distribution file in any temporary directory on your machine.

2. Log in as root on your machine.

3. From the directory containing the distribution file, run the command, for example:

   `./xtrclilite-linux-5.5`

   The exact command will vary based on your operating system and the distribution version number. If you are unable to run this command, check to make sure the distribution file has execute permission.

4. This will install OCR Shop XTR Lite, placing these files on your system:

```
/opt/Vividata/bin:
   xtrclilite
   vvlmreread
   vvlmstop
   vvlmhostid
   vvlmstatus
   gs

/opt/Vividata/bin/linux:
   xtrclilite
   vvlmreread
   vvlmstop
   vvlmhostid
   vvlmstatus

/opt/Vividata/ghostscript:
   Fontmap
   Fontmap.GS

/opt/Vividata/docs:
   xtrclilite_README.txt
   xtrclilite_release_notes.txt
```

```
   xtrclilite_quickstart.txt
   xtrclilite_manual.pdf

/opt/Vividata/docs/sample_images:
   letter.tif
   letter.pdf
   cyrillic_with_english.tif
   french_german.tif

/opt/Vividata/lib/langs:
   asciieng.lng
   BALTIC.shp
   CharSetTable.chr
   CYRILLIC.shp
   czech.lng
   danish.lng
   dutch.lng
   english.lng
   finnish.lng
   french.lng
   german.lng
   greek.lng
   GREEK.shp
   hungar.lng
   italian.lng
   LATIN1.shp
   LATIN2.shp
   norsk.lng
   polish.lng
   port.lng
   russian.lng
   spanish.lng
   swedish.lng
   turkish.lng
   TURKISH.shp
```

5.You must install the license key next before you can run OCR Shop XTR Lite.

How to install in a custom directory

By default OCR Shop XTR Lite installs in `/opt/Vividata/`. If you would like to install OCR Shop XTR Lite elsewhere, set the environment variable `VV_HOME` to the directory where you wish to install. You must set `VV_HOME` before:

- Running the installer
- Installing the license key
- Running the xtrclilite program

For example, in tcsh, to install OCR Shop XTR Lite in `/home/smith/`, set `VV_HOME` with the following command:

```
setenv VV_HOME /home/smith
```

In this manner, you may install as a non-root user, if you have write permission for the installation directory.

Temporary files

To set the temp directory used during installation and usage of OCR Shop XTR Lite, set the environment variable TMP_DIR to a specific directory. /tmp is used by default. Make sure the temp directory gives read and write permission to the user running OCR Shop XTR Lite.

Please see the section "Configuring Your Environment" for more information about setting environment variables.

About Licensing

An OCR Shop XTR Lite license key must be installed on your machine before you can run the software. The license key:

- Indicates you are authorized to run OCR Shop XTR Lite.
- Specifies an expiration date for temporary or evaluation licenses.
- Controls the number of concurrent instances of the software you may run.
- Controls which features or add-on modules you may use based on your purchase, including PDF/PS input and additional languages.

The license key is stored in the license file, by default `/opt/Vividata/config/vvlicense.dat`. If you have multiple Vividata products installed, the vvlicense.dat file will contain multiple license keys.

OCR Shop XTR Lite uses the Vividata license manager to read the license key. The license manager is started automatically the first time you run `xtrclilite`, then continues to run as a daemon process in the background. The daemon process is named “vvlicense” or “xtrclilite,” depending on your system. The license manager daemon process should be left running, unless you encounter a problem with licensing or need to install a new key.

Several license manager utilities come with OCR Shop XTR Lite:

`vvlmstop`: Shuts down the license manager process.

`vvlmreread`: Re-reads the license key; this utility should be used if you replace /opt/Vividata/config/vvlicense.dat by hand.

`vvlmhostid`: Displays the machine id of the system; the machine id is used to tie a license key to a specific machine.

`vvlmstatus`: Prints out the license manager status, including the number of instances available; this command is only meaningful if the license manager daemon has been started.

`xtrapiKeyRead`: Decodes a license key string, providing information about the machine id, expiration date for evaluations, and licensed features associated with the key.

The license manager is always started implicitly by running the normal `xtrclilite` binary.

To install your license key

You should be able to download your license key from Vividata's website after registering for an OCR Shop XTR Lite evaluation or purchasing the software. Log in at www.vividata.com using your email address and the password sent to you after registration. Go to the "License Keys" page and download your OCR Shop XTR Lite license key.

If you have trouble finding or downloading your license key, please contact Vividata.

Your license key comes wrapped in a shell script for installation. To install your license key:

1. Log in as root on your machine where OCR Shop XTR Lite is installed. (If you installed as a non-root user in a custom directory, you do not need to log in as root.)

2. If you are using a directory other than the default `/opt/Vividata/` for the Vividata installation directory, set the environment variable `VV_HOME` to your custom Vividata installation directory.

3. Run the provided license install script with the command (your license install file may have a different name):

   ```
   # sh key.sh
   ```

4. The license should now be installed on the target machine in `/opt/Vividata/config/vvlicense.dat`, or another directory as specified by `VV_HOME`. The file vvlicense.dat contains your actual license key.

   If you already had a Vividata license key installed for another product, the OCR Shop XTR Lite license key will be appended to the vvlicense.dat file, so you may continue to use all Vividata products.

Removing OCR Shop XTR Lite

Should it be necessary to remove OCR Shop XTR Lite from your system, log in as root and execute the following command:

```
# rm -r <vividata install directory>
```

Configuring the Environment

OCR Shop XTR Lite recognizes the following optional environment variables:

VV_HOME

VV_HOME specifies the location where OCR Shop XTR Lite is installed, so that `xtrclilite` can find the resource files for recognition. If VV_HOME is not set, `xtrclilite` will look for the resource files in the default installation directory, `/opt/Vividata`.

You only need to set VV_HOME if you installed OCR Shop XTR Lite in a directory other than `/opt/Vividata`.

TMP_DIR

TMP_DIR is the location where temporary files are created. By default, `/tmp` is used. This directory must have read and write permission for the user running OCR Shop XTR Lite.

You only need to set TMP_DIR if you want to store temporary files someplace other than `/tmp`.

VV_DEBUG

VV_DEBUG is a flag used to signal to the program to print out varying levels of verbose informational and debugging output. If you encounter problems with OCR Shop XTR Lite and want to report an error to the Vividata support department, we recommend setting `VV_DEBUG` to `1000`. Other meaningful values for `VV_DEBUG` are:

`1`	Only error messages (default)
`250`	Warning messages
`500`	Informational messages
`510`	More informational messages
`1000`	All messages

VV_LANGUAGE

The VV_LANGUAGE environment variable specifies one or more languages to use for recognition. If VV_LANGUAGE is not set, English is used by default. For information on how to set VV_LANGUAGE, see Chapter 3, "The language option."

VV_OUTPUTFORMAT

The VV_OUTPUTFORMAT environment variable specifies the output file format, and may be set to "unicode" or "ascii." The default is "ascii." For more information on VV_OUTPUTFORMAT, see Chapter 3, "The output format option."

Environment variables may be set from the command-line before you run xtrclilite or in your shell configuration file (.cshrc or .bashrc, for example).

If you need to set an environment variable temporarily, it is usually easiest to set it on the command-line. However, if expect to always need one of these environment variables set, then we recommend setting it in your shell configuration file. For example, if you installed OCR Shop XTR Lite in a non-default directory, then it would make sense to set `VV_HOME` in your shell configuration file.

In addition, consider adding the name of the directory that contains OCR Shop XTR Lite to the PATH

environment variable assignment in your shell configuration file. This will allow you to launch the application from any directory without specifying the full path.

After modifying your shell configuration file, logout from the system and then login again to start your session with the modified initialization.

The command used to set an environment variable depends on which shell you use. Refer to your shell documentation for instructions on how to set an environment variable.

Chapter 3: Using OCR Shop XTR Lite

Overview

The OCR Shop XTR Lite binary is `xtrclilite`, which is installed in `/opt/Vividata/bin` by default. `xtrclilite` should be invoked from the command-line or from a script for batch processing.

When you invoke `xtrclilite`, you must either use the fully qualified path such as **`/opt/Vividata/bin/xtrclilite`**, or you should add the Vividata binary directory to your `PATH` environment variable. If the Vividata binary directory, by default `/opt/Vividata/bin`, is listed in your PATH environment variable, then you may run `xtrclilite` from any directory simply by typing **`xtrclilite`** followed by the parameters, without specifing the full path.

In running OCR Shop XTR Lite, there are a number of ways to speed up recognition, increase accuracy, and streamline OCR workflow. These are covered throughout the following chapters of this manual.

Usage and options

OCR Shop XTR Lite has two special options to print command-line help and version information:

```
xtrclilite --help

xtrclilite --version
```

For normal document recognition, the OCR Shop XTR Lite syntax is:

```
xtrclilite <input filename> <output filename> [options]
```

The input filename and output filename are required. The filenames may be specified by a relative path or a full path. When xtrclilite is invoked, it processes the specified input file, generates the output file, then returns with an exit code of zero on success or a negative number on failure.

The following command-line options are available, but are not required to run OCR Shop XTR Lite:

-l <languagename,...>	Language used for recognition See list below for available values. Multiple languages may be specified and should be separated by commas. Languages not supported by ASCII text will require the output format to be set to "unicode." Default: english
-o unicode|ascii	Output format Either "unicode" or "ascii" may be specified. Languages not supported by ASCII text will require the output format to be set to "unicode." Default: ascii

In most cases, the language and output format options are specified on the command-line. However, the language and output format may also be specified by setting these environment variables:

VV_LANGUAGE	Language used for recognition Corresponds to the "-l" command-line option. May be set to one or more languages, separated by commas, in the same manner as the command-line option.
VV_OUTPUTFORMAT	Output format Corresponds to the "-o" command-line option. May be set to "unicode" or "ascii," in the same manner as the command-line option.

Environment variables are useful when a user expects to run a large number of command-lines by hand with non-default settings. By setting the environment variables, the user can avoid typing the command-line options each time.

If both a command-line option and an environment variable are set, then the command-line option will take precedence. See the section Configuring Your Environment for information on how to set environment variables.

The input file

The following image file types are supported as input:

- Multipage and single page TIFF files
- GIF (Graphics Interchange Format)
- JPEG (Joint Photographics Experts Group File Interchange Format)
- PDF (Portable document format)*
- PS (Postscript®)*
- PBM (Portable BitMap)
- PNG (Portable network Graphics Format)
- PPM (Portable PixMap)

- Rasterfile
- SGI-RGB (Silicon Graphics image file format)
- XWD
- X11

The ability to process PDF and Postscript (PS) input files is sold as an add-on module and is licensed separately. In order to process PDF and PS files, you must have a license that enables this feature. Users who purchase the PDF/PS input option and evaluation users have access to this feature.

OCR Shop XTR Lite renders PDF and PS files with Ghostscript (gs), which is distributed with OCR Shop XTR Lite.

Corrupt or invalid input files will not be able to be processed and will result in an error.

OCR Shop XTR reads black and white, grayscale, and color input images at 1-, 8-, and 24-bit depths. OCR Shop XTR Lite internally converts grayscale and color images to black and white for OCR processing.

The language option

OCR Shop XTR Lite recognizes a document with respect to one or more specific languages. The OCR engine analyzes the document's character set and, when a dictionary is available, words based on the specified language or languages.

56 languages are available for use with OCR Shop XTR Lite. By default, the English language is used. When a document is written in a language other than English, the user should set the "-l" language option or the VV_LANGUAGE environment variable to specify the language of the document. Having the engine recognize the document with respect to the correct language greatly improves results. The command-line option value will override the environment variable value, if both are set.

OCR Shop XTR Lite reports the language(s) used during recognition to the console, so it is clear to the user which languages were selected.

Some documents contain text in multiple languages. OCR Shop XTR Lite allows the user to specify more than one language at a time, separated by commas. Multiple languages must come from the same code page, as outlined in the tables below. The one exception is "asciieng", which may be specified alongside any other language, thus allowing Latin 1 characters to be recognized alongside non-Latin 1 characters.

Some languages are not supported by ASCII text, as outlined in the table below. When using these languages, the user must select the Unicode output format.

It is important to select the correct language for a document, because OCR Shop XTR Lite uses information about that language, including which characters or glyphs are used, when performing OCR. In addition, some languages come with dictionaries, which further improves OCR results by contributing word information.

The languages supported by OCR Shop XTR Lite are divided into six different code pages according to the shapes or glyphs of characters contained within the language. The OCR engine follows the glyph conventions of Microsoft Code Pages. See http://msdn.microsoft.com/en-us/goglobal/bb964653 for more details. If you are recognizing multiple languages at a time, it is important to know whether they belong to

the same code page or not, as outlined in the table below. Only languages that belong to the same code page may be recognized at the same time. “asciieng” is an exception and may be used alongside any other language to allow recognition of English characters.

Valid language names for the “-l” command-line option and the `VV_LANGUAGE` environment variable are:

afrikaans	friulian	polish
albanian	gaelic	port (Portuguese)
aymara	galician	romanian
basque	german	russian
breton	greek	serbian
bulgarian	greenlandic	sbcroatian (Serbo-Croatian)
byelorussian	hawaiian	slovak
catalan	hungar (Hungarian)	slovenian
croatian	icelandic	sorbianl (Sorbian – Lower)
czech	indonesian	sorbianu (Sorbian – Upper)
danish	italian	spanish
dutch	kurdishlat	swahili
english	latin	swedish
estonian	latvian	tahitian
faroese	lithuanian	turkish
finnish	macedonianc	ukranian
flemish	malaysian	welsh
french	norsk	zulu
frisianw (Frisian – West)	piginenglish	asciieng

The languages grouped by code page are:

Baltic (1257):	**Central Europe (1250) :**	**Cyrillic (1251):**	**Greek (1253):**	**Turkish (1254):**
Estonian	Albanian	Bulgarian	Greek	Kurdish
Latvian	Polish	Macedonian		Turkish
Hawaiian	Croatian	Serbian		
Lithuanian	Romanian	Byelorussian		
	Slovenian	Russian		
	Czech	Ukranian		
	Serbo-Croatian			
	Sorbian - Lower			
	Hungarian			
	Slovak			
	Sorbian – Upper			

Latin I (1252)		
Afrikaans	Portuguese	Icelandic
French	Catalan	Tahitian
Malaysian	Galician	Faroese
Aymara	Spanish	Indonesian
Frisian - West	Danish	Welsh
Norwegian	German	Finnish
Basque	Swahili	Italian
Friulian	Dutch	Zulu
Pigin English	Greenlandic	Flemish
Breton	Swedish	Latin
Gaelic	English	

The languages grouped by supported output format are:

These languages may use the ASCII or Unicode output formats:	These languages must use the Unicode output format:
afrikaans aymara basque breton catalan danish dutch english faroese finnish flemish french frisianw (Frisian – West) friulian gaelic galician german greenlandic icelandic indonesian italian latin malaysian norsk piginenglish port (Portuguese) spanish swahili swedish tahitian welsh zulu	albanian bulgarian byelorussian croatian czech estonian greek hawaiian hungar (Hungarian) kurdishlat latvian lithuanian macedonianc polish romanian russian serbian sbcroatian (Serbo-Croatian) slovak slovenian sorbianl (Sorbian – Lower) sorbianu (Sorbian – Upper) turkish ukranian

The output format option

OCR Shop XTR Lite offers two types of text output: ASCII and Unicode. ASCII text is used by default. The user may specify which output format to use on the command-line with the "-o" option or by using the VV_OUTPUTFORMAT environment variable.

While the Unicode format supports all languages, ASCII text does not. Documents recognized in a language not supported by ASCII text must use Unicode as the output format. Refer to the table above for a list of which languages are supported by both ASCII and Unicode text, and which languages are only

supported by Unicode text..

Valid values for the "-o" option and `VV_OUTPUTFORMAT` are:

ascii
unicode

When Unicode text is generated, the output file must be viewed in a Unicode-compatible viewer. Refer to www.unicode.org for more information on the Unicode format and viewing Unicode text files.

The return code

When OCR Shop XTR Lite completes, an exit code of 0 is returned on success. If there was an error or OCR Shop XTR Lite was unable to complete for any reason, a negative number is returned to indicate failure.

Chapter 4: Examples

Overview

In order to facilitate use of OCR Shop XTR Lite, this section presents a series of example commands to be issued via the command line.

Examples throughout this manual assume that your `PATH` environment variable has been set to include the Vividata binary directory, `/opt/Vividata/bin by default`. If it has not, you should specify the full path when running `xtrclilite`.

OCR Shop XTR Lite is a command-line interface program and should be invoked on the command-line or from a script in the same way as other applications.

All sample input files mentioned in these examples are included in the OCR Shop XTR Lite distribution and are installed in the directory `/opt/Vividata/docs/sample_images`. Each command-line in these examples should be typed as one line; in this manual, the margins cause some command-line examples to wrap, so continued lines are indicated with a backslash.

Help and version information

To print help information to the command-line:

```
xtrclilite --help
```

To print version information:

```
xtrclilite --version
```

Basic document recognition

The simplest command-line to recognize a document is:

```
xtrclilite letter.tif out.txt
```

where image.tif is the input document image. After processing image.tif, OCR Shop XTR Lite will create an ASCII output file called out.txt, which you should see in your directory when `xtrclilite` returns. View the output text file and verify that image.tif was recognized.

This first example recognizes text in English, the default. The next example explores recognition in additional languages.

Setting the language

To recognize an input document in French and generate ASCII output:

```
xtrclilite french_german.tif out.txt –l french
```

Look at the messages printed to the console to verify that the French language was used during recognition.

To recognize an input document in German and French and generate ASCII output:

```
xtrclilite french_german.tif out.txt –l german,french
```

Make sure that multiple languages are separated by commas only (no spaces).

If multiple languages are listed, they must come from the same code page (refer to the tables above). For example, Polish and Czech may be recognized together, because they both use the Central Europe code page (1250). However, Russian and French cannot be recognized together, because Russian uses the Cyrillic code page (1251) and French uses the Latin 1 code page (1252).

The “asciieng” language is an exception, because it may be specified with any other language. The “asciieng” language should be used when Latin 1 characters, such as those found in English, need to be recognized alongside one or more non-Latin 1 languages. For example, if a document contains Russian and English text, you could recognize both Russian and English characters with the command-line:

```
xtrclilite cyrillic_with_english.tif out.txt –l russian,asciieng \
–o unicode
```

The output file will contain text in both English characters and Russian characters, although the English dictionary will not be used. If “asciieng” were not specified in this case, then the OCR engine would have tried to interpret the English characters as Russian, with poor results.

The “asciieng” language option is appropriate to use in any case where a document contains Latin 1 characters alongside those from another character set. For instance, “-l greek,asciieng” is the correct setting for a document that contains both Greek and German characters.

The output file in this example must be in the Unicode format to contain Russian characters. The next section further explores how to set the output format.

Setting the output format

By default, output files are generated as ASCII text. Alternatively, the user may choose to generate Unicode text as output with the “-o” option:

```
xtrclilite letter.tif out.txt –o unicode
```

Some languages cannot be represented by ASCII text. In these cases, the Unicode output format is required. For example, to recognize a Russian document, Unicode output must be selected:

```
xtrclilite cyrillic_with_english.tif out.txt –l russian –o unicode
```

To recognize an input document in Russian and Latin 1 characters, and generate Unicode output:

```
xtrclilite cyrillic_with_english.tif out.txt –l russian,asciieng \
–o unicode
```

When viewing the Unicode output files, you must use a viewer that supports the Unicode format. Refer to www.unicode.org for more information on the Unicode format and how to view Unicode files.

Using PDF and Postscript input files

PDF and Postscript (PS) input files are detected automatically. While these formats are licensed as add-on modules, they are read in by xtrclilite in the same manner as other input files.

For example, to recognize a PDF input file and generate ASCII text output using the English language, run the command:

```
xtrclilite letter.pdf out.txt
```

Chapter 5: The Recognition Process

Overview

What Is Optical Character Recognition (OCR)?

Optical Character Recognition (OCR) is the process of converting a text image file into a text file. OCR is also referred to as text or page recognition software as it `recognizes' imaged characters and turns them into type. OCR Shop XTR Lite uses as its source document images stored in files.

OCR Shop XTR Lite begins with an image that is really just a `picture' of text and graphics and which cannot be edited directly by the user. The process of OCR transforms this `picture' into separate characters of text.

The recognized text from OCR Shop XTR Lite is then exported to an ASCII or Unicode text file.

OCR Shop XTR Lite operates in a four-step process:

1. Acquire an image from a file
2. Perform pre-processing and segmentation of the image into regions
3. Recognize the image based on a specific language or languages
4. Output the results as ASCII or Unicode text

Image input

The OCR process begins when OCR Shop XTR Lite reads in the input image data from the file specified on the command-line. OCR Shop XTR Lite automatically detects and supports a large number of input image file formats, listed below. Next, OCR Shop XTR Lite internally converts the image data into 1-bit, black and white image data, if needed, prior to running the pre-processing, segmentation, and recognition steps.

OCR Shop XTR Lite supports single and multi-page input documents. When a multi-page input image is submitted, OCR Shop XTR Lite will create one output text file that contains the recognized text from every page of the input file.

Support for PDF and Postscript input files is licensed and must be purchased as an add-on module.

The supported input image formats are:

Format name	Typical file extension	Comments
GIF (Graphics Interchange Format)	.gif	
JPEG (Joint Photographics Experts Group File Interchange Format)	.jpg	JPEG formats compatible with libjpeg, in particular JPEG File Interchange Format (JFIF). See www.ijg.org.
PBM (Portable BitMap)	.pbm	Bi-level (1-bit per pixel) bitmap with a simple header
PDF (Portable document format)	.pdf	Licensed as an add-on module.
PNG (Portable network graphics format)	.png	
PPM (Portable PixMap)	.ppm	RGB-encoded bitmap with a simple header
PS (Postscript)	.ps, .eps	Levels 1, 2, and 3 Licensed as an add-on module.
Rasterfile	.ras	Native bitmap of Sun™ ; see /usr/include/rasterfile.h or man rasterfile(5) on a Sun™ system.
SGI-RGB (Silicone Graphics image file format)	.rgb, .sgi, .iris	As per Silicon Graphics' library libimage.a. Type "man 4 rgb" on an IRIX machine for more details.
TIFF (Tagged image file format)	.tif, .tiff	Both single page and multipage formats are supported. Images within a TIFF file may be uncompressed or compressed in Group 3, Group 4, LZW, or JPEG compression.
XWD (X Window Dump)	.xwd	Screen dump from an X Window System. See “man xwd” on most Unix and Linux™ systems for details.
X11	.xpm, .xbm, .bm	X Consortium created format. It is a relatively simple bitmap format. The extension of .xpm indicates a pixel map (see "Portable PixMap (PPM)". above) and .xbm indicates a bi-level bitmap (see "Portable BitMap (PBM)" above).

Pre-processing and recognition

After reading in the input image data, OCR Shop XTR operates in fully-automatic mode to pre-process and recognize the document. Features include: auto-detection of rotated images, algorithms to detect and correct image skew, automatic filtering for fax and dot-matrix documents, analysis of the layout to segment textual and graphical regions, analysis of tables and forms to generate coherent ordering of the output text.

Font point-size may range from 5 – 72 points. Supported input image resolutions range from 72 dpi – 900 dpi.

Text output

After recognition completes, OCR Shop XTR Lite produces text output, creating a file in either ASCII or Unicode text format, as specified by the user. One output text file is created for the one input document. If the input document contains multiple pages, the output text file will contain text for every page.

If the user calls OCR Shop XTR Lite with an output filename of an existing file, the existing file will be overwritten with the new output file.

In the output text file, unrecognized characters are represented by a tilde mark: ~

Chapter 6: Analyzing and Improving Results

Overview

Typeset, high-quality printed pages return the best recognition accuracy. This chapter describes the most common factors that affect recognition quality and suggest ways of handling trouble images and improving OCR results.

OCR Shop XTR Lite is intended to handle most input images. However, when the quality of the input images is poor, varies widely, or requires adjustment, the full OCR Shop XTR product is a better choice than OCR Shop XTR Lite. The full OCR Shop XTR offers a large number of options to compensate for lower quality images and improve recognition results. See the section, "When to use OCR Shop XTR instead of OCR Shop XTR Lite," at the end of this chapter.

Analyzing document quality

High quality input documents produce the best OCR results. The quality of an input document may be analyzed based on these qualities:

- High contrast images produce the best results. Grayscale and color images where the font shade is similar to the background shade may not convert to black-and-white image data well. This is not an issue for 1-bit, black-and-white input images.

- Clean and crisp print, where characters are well-formed, distinct, separated from each other and not overlapping will produce the best results.

- Extraneous spaces within a character, overlapping characters, and poorly formed characters can result in slower or less accurate recognition. For instance, if the left bar of an "m" is not connected to the first curve, the character might be recognized as the sequence "in" instead of "m."

- Very small or very large print can be harder for the OCR engine to recognize. Analyze the size of the print with the document resolution (dpi) taken into account. View the actual pixels to evaluate how much information is available for very small characters.

- Handwritten notes, lines and doodles can slow recognition and distort printed characters. Documents that contain a large amount of background noise as a result of dithering or many small marks as in a background pattern will result in slower recognition and a higher error rate.

- While a large variety of fonts are supported, highly stylized fonts may not be recognized well.

- Underlines change the shape of descenders on the letters q, g, y, p, and j.

- Text ideally appears over a background of a solid color. Documents that contain text that overlays an image, such as labels on a map, will take a long time to process and will have a lower

recognition rate.

- Skewed images take longer to process as the OCR engine corrects the skew.

Scanning recommendations

If you have control over the scanning process, you can improve recognition by taking a few steps during scanning to eliminate skew and background noise.

- Make sure that the document is positioned correctly in your scanner and is not slanted. Words that are cut off by the scanner will be lost. OCR Shop XTR Lite automatically corrects for skew, but if a skewed image can be avoided in the first place, OCR will run faster and the results will be more accurate.

- The sheet of glass on the flatbed of the scanner must be clean and clear. If it gets dirty, wipe it gently with a soft, damp, lint-free cloth or tissue. Be sure it is completely dry before you place anything on it. Some paper is so thin that the scanner reads text printed on the back side of the scanned page. This is often the case with telephone book pages. To correct this problem, put a black piece of paper between the sheet and the lid of the scanner.

- While OCR Shop XTR Lite recognizes images with resolutions from 72 dpi to 900 dpi, we recommend scanning at 300 dpi.

- OCR Shop XTR Lite runs OCR on black-and-white image data. Unless you need color or grayscale images for another purpose, we recommend scanning at a bit depth of 1 to create black-and-white images for input to OCR Shop XTR Lite.

- Scan several typical images as a test and adjust your scanning options, if needed, to obtain clear, well-defined text.

- If you scan and create grayscale or color images, make sure that the contrast is high so that when OCR Shop XTR Lite converts them images to 1-bit black-and-white image data, enough detail is retained. If needed, adjust the contrast for your scanner to provide better input to OCR Shop XTR Lite.

About line art and photographic regions

OCR Shop XTR Lite may recognize some line-art graphics or areas of photographic regions as text if the artwork is poor and the lines resemble letter strokes or features look like in text in a light background.

About font size and resolution

Make sure the resolution of the input image, as well as the font size with respect to that resolution, are

within normal limits.

OCR Shop XTR Lite accepts:

- Image resolutions from 72dpi to 900dpi
- Fonts from 5 to 72 points

Resolutions and font sizes that fall within the middle of the range are more easily recognized and result in more accurate output than resolutions and font sizes that are either very small or very large.

The resolution of the input image determines what one "point" means in the font point size. The resolution of the input image is specified in the input image file, or, when not specified, is assumed to be 300x300 dpi by default.

- There are approximately 72 points per inch.
- The point size of a font is measured from the top of the highest ascender to the bottom of the lowest descender.
- The dpi specifies the number of pixels per inch.

If the type in your image is particularly large or small, it might fall outside the accepted font point sizes, depending on the image resolution.

For example, if your font size is 15 pixels high and the image resolution is 300dpi, then the font point size is approximately 3 points, too small for the engine to recognize well. Similarly, if your font size is 80 pixels high and the image resolution is 72dpi, then the font point size is approximately 80 points, too large for the engine to recognize well.

You may approximate the point size of your font with the equation:

[height of font in pixels] * 72 points/inch / [image dpi] = [point size of font]

Remember the height of the font is measured from the top of the highest ascender to the bottom of the lowest descender. If you count the pixels, make sure you view that portion of your image at full resolution on your screen, sometimes referred to as "actual pixels".

What you can do:

If you suspect the resolution of your input image is causing the fonts to fall outside the accepted range, or if your input image resolution is larger or smaller than the supported resolutions, adjust the settings on your scanner and scan your images at a supported resolution. A good starting point is 300x300 dpi.

If you need more flexibility, Vividata's full OCR Shop XTR software includes a command-line option to override the input image resolution. Overriding the input image resolution lets you recognize a wider variety of input images with unusual resolutions or resolutions outside of the supported range. Without the full OCR Shop XTR, if you do not have control over the scanning process, you will need to preprocess your image yourself. Try using an image processing application to change the resolution of your input image before recognizing it with OCR Shop XTR Lite.

About image contrast

In low contrast images, the text is not significantly darker than the background color. Sometimes this can cause recognition problems when OCR Shop XTR Lite converts the image data to black-and-white, 1-bit image data prior to recognition. Information might be lost during the conversion when the input image is very low in contrast.

What to do:

If you have control over your scanning process, increase the contrast when you scan your images.

Alternatively, use a third party image processing application to increase the contrast in your image files before passing them to OCR Shop XTR Lite.

If you need more flexibility, try Vividata's OCR Shop XTR software, where one of the command-line options permits you to adjust the threshold at which input images are binarized. An evaluation of OCR Shop XTR is available from Vividata's website, or please contact Vividata directly.

When to use OCR Shop XTR instead of OCR Shop XTR Lite

OCR Shop XTR Lite works well for applications where standard input documents need to be converted into text output files. However, the full OCR Shop XTR works better for applications that require more flexibility and customization. Go to Vividata's website www.vividata.com to read more about OCR Shop XTR and to download a software evaluation. Vividata Support or Sales can also provide information to help you decide which software product will work best with your application.

Here are situations where OCR Shop XTR is a better choice than OCR Shop XTR Lite:

- When output is needed in a format other than text: Besides ASCII and Unicode text, OCR Shop XTR offers PDF, HTML, XDOC, and 8-bit output.

- Forms processing: OCR Shop XTR permits the user to set up custom regions to batch process forms and identify fields.

- Low-contrast input images: OCR Shop XTR provides more a tool for adjusting how the input image is rendered, allowing the user to compensate for low-contrast input images.

- Resolution control: OCR Shop XTR lets the user override the input image resolution. This is useful when input images do not contain resolution information, when they contain incorrect resolution information, and when the interpreted font size needs to be manipulated so that it falls within the support font size range.

- Confidence values: OCR Shop XTR's XDOC output provides information on word and/or character confidence values, giving the user feedback on the recognition results.

- Page formatting and word location: OCR Shop XTR's XDOC output can provide information on

the placement of words, spaces, images, and paragraphs, as well as information about fonts and page formatting.

- Graphics output: OCR Shop XTR can generate output image files of both image and text regions.

- Character set restriction: OCR Shop XTR permits the user to restrict the character set so that only a subset of all characters are recognized, thus improving results on input documents where a subset of characters is used.

- More control: OCR Shop XTR lets the user choose which filters to apply, when to analyze the image, and how to process the image data when automatic processing is not sufficient.

- More input/output flexibility: Multiple input files may be passed to OCR Shop XTR to create multiple output files or a single output file that includes OCR results from all passed input files.

- More efficiency: Unneeded automatic processing features may be turned off when the user knows they will not be needed, thus speeding up processing time.

Appendix 1: Troubleshooting

Overview

This chapter offers some troubleshooting hints as well as brief pointers to maximize operation efficiency.

Technical Support Questions

View the Support and Documentation sections of our website: www.vividata.com. Send questions to Vividata Technical Support by clicking Contact Support.

Identifying the Problem

Below are several steps to help you identify the cause if OCR Shop XTR Lite fails to run. The following commands assume that the Vividata binary directory, unless installed elsewhere `/opt/Vividata/bin`, is in your path.

Is licensing working?

A. Make sure the license manager is running. It is started when you first run **`xtrclilite`** but is shut down when you either reboot your machine or run **`vvlmstop`**.
B. Run **`vvlmstatus`**
C. It should list your installed license keys and how many are available. Your license controls how many concurrent instances of OCR Shop XTR Lite you may run and which features are enabled.
D. If the license status looks ok, then go to step 2. If you suspect a problem with the license manager, refer to Appendix 2.

Are you unable to run the `xtrclilite` binary?

Verify that the Vividata installation directory is included in your `PATH` environment variable or that you are calling **`xtrclilite`** using an absolute path.

Are library files (language files and character set tables) not found?

If you installed OCR Shop XTR Lite in a directory other than the default `/opt/Vividata`, make sure the `VV_HOME` environment variable is set.

Is there a problem reading the input image file?

Make sure your input file exists in the location you specify on the **`xtrclilite`** command line, and that it has read permission.

Verify that the input file is not corrupt by opening it with an image viewer or appropriate program.

If the input file is either a PDF or Postscript (PS) file, verify that you are licensed for PDF/PS input. If your license does not include the PDF/PS input feature, you will not be able to process PDF or Postscript input files. Evaluation licenses typically include the PDF/PS input feature. Purchased licenses include the PDF/PS input feature if you purchased the PDF/PS input add-on module or feature.

If the input file is a PDF or Postscript file and your license supports the PDF/PS input feature, verify that the input file is a valid PDF or Postscript file by opening it with Ghostscript. For example, use Ghostscript to open the file on the screen by running the command: **`/opt/Vividata/bin/gs image.pdf`**

Is there a problem creating the output file?

Verify that the output file is written to a directory with write permission. If a file already exists of the same name as the desired output file, make sure you have write permission to it.

Do you have write permission for the temporary directory?

Verify that you have read and write permission to the temporary directory, either `/tmp` or the directory specified by the environment variable `TMP_DIR`. OCR Shop XTR Lite, Ghostscript, and the license manager all might use the temporary directory.

If any Vividata or Ghostscript logs already exist in the temporary directory, verify that you have read and write permission to those files.

Are the OCR results poor?

If OCR Shop XTR Lite generates poorer results than you expect, here are a few things to check:

1. What language is used in your input document?

   OCR is performed for the English language by default. If your input document is in another language, make sure you select the correct language or languages using the "`-l`" command-line option for the `VV_LANGUAGE` environment variable.

   You can verify which language was used during recognition by viewing the informational messages printed to the console when you run `xtrclilite`.

2. What is the resolution of the input file?

   The OCR engine supports resolutions of 72 dpi to 900 dpi.

   If your input file's resolution falls outside the supported range, the OCR engine will generally not be able to generate results as good as what otherwise may be achieved.

If the resolution is on the edge of this range, especially if the fonts are very small or very large, then the results will not be as good as for images with resolutions in the middle of the range. Read the section “Font size and recognition” in Chapter 6.

Consider using OCR Shop XTR instead of OCR Shop XTR Lite. OCR Shop XTR offers options for manipulating the image resolution and thus improving OCR results.

3. Is the input image of a high quality?

   View your image in a graphical image viewer. Are the letters clear and distinct? Is there space between each letter? Are the shapes of the characters distinct? Is there handwriting or extraneous marks obscuring portions of the type?

   View the image at full resolution, so you can see the actual pixel size of the letters. How much information is really there?

   Answers to these questions will help you estimate what kind of results to expect from an OCR engine.

4. Is the input image a low contrast image?

   If the input image is a grayscale or color image, check whether it is low contrast. Low contrast images have text that is similar in color to the background. For instance, dark gray text over a light gray background.

   High contrast images are black text on a white background, for instance, or white text on a black background.

   OCR Shop XTR Lite converts the input image to 1-bit black and white image data prior to recognition. During this process, low contrast image might have too much information lost.

   If your image is low contrast, consider using OCR Shop XTR instead of OCR Shop XTR Lite, because OCR Shop XTR offers an option for controlling how the conversion to black and white image data takes place, so that low contrast images can produce good OCR results. Alternatively, consider processing your input image with a third party image processing tool to increase the contrast.

If OCR Shop XTR Lite is still not functioning properly or your results are not what you expect, please contact Vividata's customer support.

Appendix 2: License Manager Troubleshooting

How to reset the license manager

The following section assumes the Vividata binary directory is listed in your PATH environment variable. If it is not, you will need to use the fully qualified path when invoking the Vividata binaries.

If the license manager enters a bad state, reset it by running the command:

```
vvlmstop
```

Then try running `xtrclilite` again.

Steps to handle a licensing error

If you still encounter license manager problems, check the following:

1. Is your license key installed?

   Make sure the file `/opt/Vividata/config/vvlicense.dat` exists, contains your license key, and has read permission.

   If the license file does not exist, follow the instructions in Chapter 2 to install your license key.

2. Does the machine ID of your computer match the machine ID of your license key?

   The machine ID of your computer and the machine ID encoded in your license key must match for the software to run. Obtain your machine id by running:

   `/opt/Vividata/bin/vvlmhostid`

   You may obtain the machine ID associated with your license key by either logging into your account at www.vividata.com and viewing your License Keys or you may decode the license key installed on your machine by running:

   `/opt/Vividata/bin/xtrapiKeyRead -k <keystring>`

   where <keystring> is the encoded string contained in your license file, `/opt/Vividata/config/vvlicense.dat`.

   Look at the `xtrcliKeyRead` output and check that the machine id matches that printed out when you ran vvlmhostid. Verify that your license key has not expired.

3. Are you currently evaluating OCR Shop XTR Lite?

   If so, your evaluation license may have expired. Log into your Vividata account at www.vividata.com and check when your license expires. If you need more time to evaluate, you may request an extension from Vividata Sales.

4. Is OCR Shop XTR Lite installed in a directory other than `/opt/Vividata`?

   If so, verify that the `VV_HOME` environment variable is set.

   If `VV_HOME` is not set or you suspect it may not have been set before running OCR Shop XTR for the first time, run **`vvlmstop`** to stop the license manager. Next, set `VV_HOME` to the directory where you installed OCR Shop XTR Lite; the command you use will depend on your shell. Start up the license manager again by running `xtrclilite`.

5. Are you running too many instances of xtrclilite or are you trying to use unlicensed features?

   The license manager controls how many concurrent instances of xtrclilite may run at once, as well as which features are enabled. Each invocation of xtrclilite is one instance; if you run xtrclilite in two separate shells at the same time, two concurrent instances are used. Licensed features include PDF/PS input documents and language usage.

   Evaluation license keys typically permit all features and one concurrent instance. Purchased license keys reflect the number licenses purchased and the features purchased.

   Log into your account at www.vividata.com to check how many licenses of OCR Shop XTR Lite you have and what features you purchased. Verify that you are not trying to use features for which you are not licensed or trying to run more concurrent instances than you have licenses.

6. Does the license manager log have write permission?

   Make sure the license manager log, /tmp/vvLicense.log, has read and write permission. If it is owned by root or does not have read and write permissions, delete it, run **`vvlmstop`** to stop the license manager, and run `xtrclilite` again to restart the license manager.

   Make sure you run `xtrclilite` as a normal user and not root. This problem usually arises when `xtrclilite` is initially run as root, so the license manager is started as root and root owns the license manager log. Subsequent users can then not overwrite the log owned by root.

7. Does the temporary directory have read and write permissions?

   Check that /tmp or your custom temporary directory has read and write permissions, so that the license manager can write files to it. If the environment variable `TMP_DIR` is set, OCR Shop XTR Lite and the license manager will place temporary files in the specified directory.

8. What is the license manager status?

   If the license manager is already running, you can check the license manager status with the command:

`vvlmstatus`

It should list your installed license keys and how many are available. Verify that you have the number of licenses you expect and that those licenses are available (not checked out).

If you need to start the license manager, run `xtrclilite`, which starts the license manager process in the background automatically.

How to download your license key and access your license information online

Your license keys and information about them are available from Vividata's website. To download your license key, view the features and concurrent instances enabled, or to view the machine ID of your license key:

1. Log into your Vividata account at www.vividata.com using your email address and the password sent to you after registration.

   If you need help with your user ID and password, contact Vividata Support.

2. Click on the link to view your license keys.

3. Find your current OCR Shop XTR Lite license key in the list. Verify that the machine id matches that of the computer where you installed the license, and that the features and number of concurrent instances are what you expect.

4. Contact Vividata if you have any questions about the features enabled by your license key or if you would like to move the license to a new machine.

For any other questions regarding your license key or licensing in general, please contact Vividata Support.

Appendix 3: Glossary

ADF: Automatic document feeder

ASCII: An acronym for American Standard Code for Information Interchange; a code in which the numbers from 0 to 127 stand for text characters. ASCII code is used for representing text inside a computer and for transmitting text between computers or between a computer and a peripheral device.

auto segmentation: The process in which the OCR Shop XTR Lite determines where on a page different elements are such as where pictures are and where columns of text are.

binary image: An image that is represented using only one bit per pixel. Such images are also called black and white, monochrome, bi-level, or 1-bit.

bit-mapped image: A collection of bits (dots) in memory that represent the scanned image. The display on the screen is a visible bit-mapped image.

Code Page: "Code Page" is a Microsoft® term. A code page is a particular mapping of a set of unsigned bytes to a set of visible characters (and space characters). Different code pages are used to represent in memory the characters in different languages. For more details, see: http://www.microsoft.com/globaldev/reference/WinCP.asp

digital image: A digital image is the way a picture or visual image of some object is represented in computer memory. A digital image consists of a number of pixels and a description of how the pixels are arranged to form the image. In addition, information about how each pixel stores the color of the original image is included.

Document: A document is a set of pages that are related usually because the sense of the text on one page flows into the next as in a book. For OCR Shop XTR Lite, it is best to arrange for documents to be sets of pages that have the same font or set of fonts continuing from one page to the next. This best takes advantage of the internal font learning system that is built into the OCR Shop XTR Lite recognition system.

Dpi: An abbreviation for dots per inch. This is the number of dots per linear inch that a printer can print or a scanner can produce. See also resolution.

frame: A frame is a way to represent the maximum extent of some page element in the horizontal and vertical direction (X and Y coordinates respectively). A frame can be thought of as a rectangle that is lined up with the X- and Y-axes. Frames are represented by four numbers, which can be top, left, bottom, right or top, left, height, width. Also see UOR.

language pack: A language pack is a data file supplied with the OCR Shop XTR that includes information about how the characters of a given language are put together to write words and sentences in the language. Language packs contain information about the common words used in a language, rules for punctuation and the conventions used when writing things such as numbers, money amounts and dates.

language set: A language set is that set of supported languages that can be recognized with a given shape pack loaded. Each of the supported languages in a language set may or may not have an available language pack associated with it. Languages without an available language pack can still be recognized but accuracy for these languages will not be as high as for languages for which a pack exists

monospaced font: Any font in which all characters have the same width. For example, in Courier New (a monospaced font), the letter "M" is the same width as the letter "l". Thus, "MMMMM" is the same width as "lllll".

orient: To orient a page is to rotate the page in memory so that it is better positioned for display to the user and/or recognition by the OCR Engine. A page is oriented for recognition when the text flows left to right (from low X to high X coordinates) and from top to bottom (low Y to high Y coordinates).

page: A page is the unit that makes up a document. Within OCR Shop XTR Lite, a page is usually the representation of one side of a single piece of paper if that was input from a scanner. In addition it may be a single image, input from a file, fax machine, digital camera or other digital image input device. For purposes of licensing, a page size equivalent to a US letter size or ISO A4 is used.

pixel: Pixel is short for picture element. It is a point or dot on the graphics screen. It is the smallest definable unit of a digital image. Each pixel represents a single point in the image. The number of pixels per unit distance (dot-per-inch or DPI for instance) within a digital image is referred to as the resolution of the image. A pixel can be binary, gray, or color, or can be an index into a palette. Binary pixels require only one binary digit or bit of computer memory to store; gray, color and indexed pixels use more bits with 4, 8, and 24 being common values for the number of bits used.

point: A typographic unit of measurement equal to 1/72 inch, measured vertically. Points are used to describe font size.

proportional font: Any font in which characters differ in width. For example, in the proportional font used here, the letter "M" is wider than the letter "l". Thus, "MMMMM" is wider than "lllll."

recognize: In the context of the OCR Shop XTR Lite, when an image is recognized, it is processed using the OCR Engine that is part of the OCR Shop XTR Lite. During this process the pixels making up a digital image are processed by the OCR engine to determine which pixels are parts of visible text characters within the image. The identities of those characters are also determined and stored in memory using the code page representation of the given character. The result of recognition is used to create output based on the user settings.

region: A region is an area of a page that usually contains either all text or all pictures. Regions can be determined by the OCR Shop XTR Lite during auto-segmentation or specified by a user in an rdiff input file. Regions on a page can overlap. Regions can be simple rectangles in shape or they can be more complex (see UOR).

resolution: The fineness with which a scanner, printer, or other device produces information. It is expressed in dots per inch (dpi). A higher dpi produces a sharper image.

shape pack: A shape pack is a data file supplied with the OCR Shop XTR Lite that describes the shapes of the characters that can be recognized by the OCR engine when that shape pack is loaded. Each shape pack corresponds to a particular code page that will be used for output when that shape pack is loaded. For each shape pack there is an implied language set that represents the supported languages that can be recognized with that shape pack loaded.

Skew: Skew is the amount of tilt in an input image. Skew is generally used to describe the tilt in images including text. In such images the tilt is more apparent and affects recognition and layout analysis.

swap file: An area of the hard disk that is used for temporary data storage when RAM is low or used up.

This is also known as virtual memory. A swap file lets you run more programs than you could with actual memory, but it is slower than using regular memory.

text file: A file containing information in text form; its contents are interpreted as characters encoded using the ASCII (or comparable) format.

TIFF: An abbreviation for tagged image file format. This is a standard graphic file format for grayscale and high-resolution bit-mapped images.

TrueType fonts: One of the major types of scalable fonts. These can be printed or displayed on the screen at any size.

Unicode: UNICODE is a standard for representing visible characters using a stream of bytes in computer memory or on some other digital storage medium. Unlike code pages where each code page can only be used to describe a subset of the known written languages, Unicode is a single standard way to represent all of the world's common written languages. Whereas the code page representation uses a single byte to represent each character, Unicode uses a 16-bit word for each character. The OCR engine that is part of OCR Shop XTR Lite does recognition internally based on a single selected code page. During output however, the text data can be converted to Unicode for use with other applications that expect text data in Unicode format.

zone: See Region.