Simple Zonal OCR Manual

Revised 9-18-2011

eDocFile, Inc. ,2709 Willow OaKs Drive, Valrico, FL 33594 - Phone 813-413-5599
eMail sales@edocfile.com Website www.edocfile.com

Simple Zonal OCR - Purpose

Simple Zonal OCR is a program that monitors a file folder for Tiff images. When one is found, if it is a multi-page tiff image it can be split into separate documents if there is a blank page separator being used. Once split, up to two areas can be selected to be processed. The processing includes OCR, applying rules for text extraction to the OCR for validation, and renaming the file as a Tiff or PDF with the extracted data.

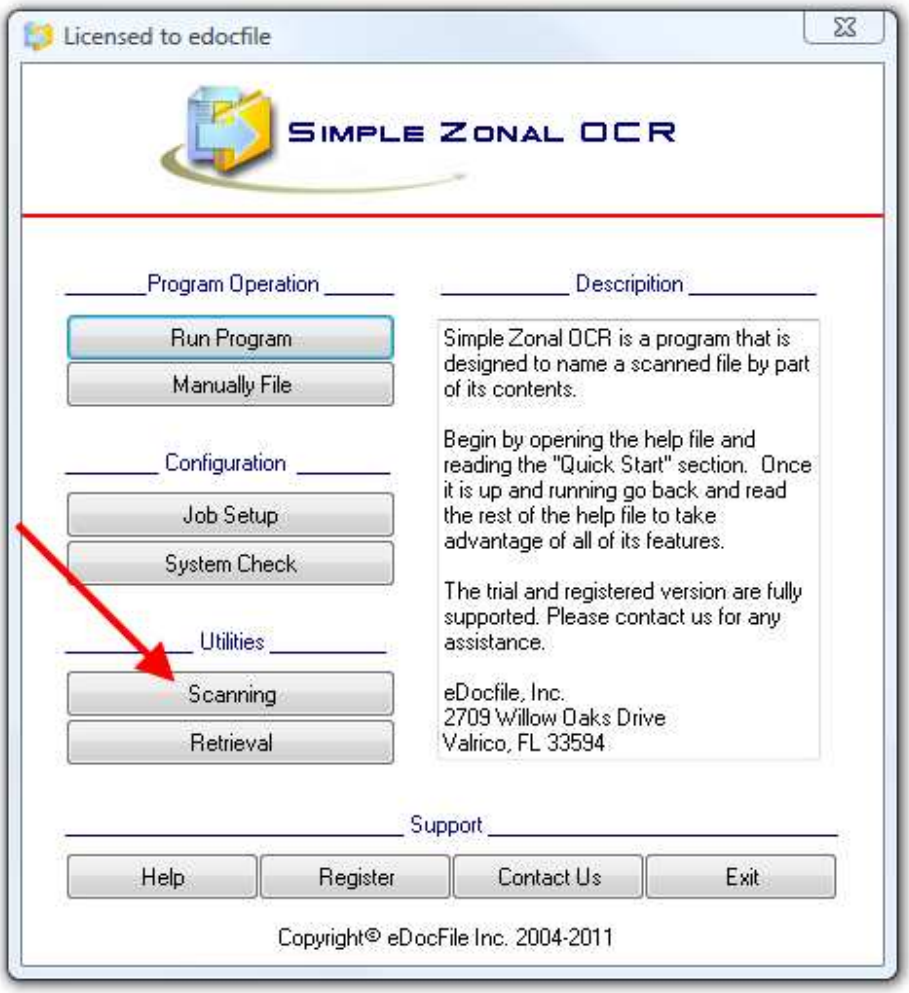
It is ideally suited for processing Work Orders, Delivery Tickets, and all other documents that are generated that can be referenced via a number or text string that is always located in the same area.

System Requirements

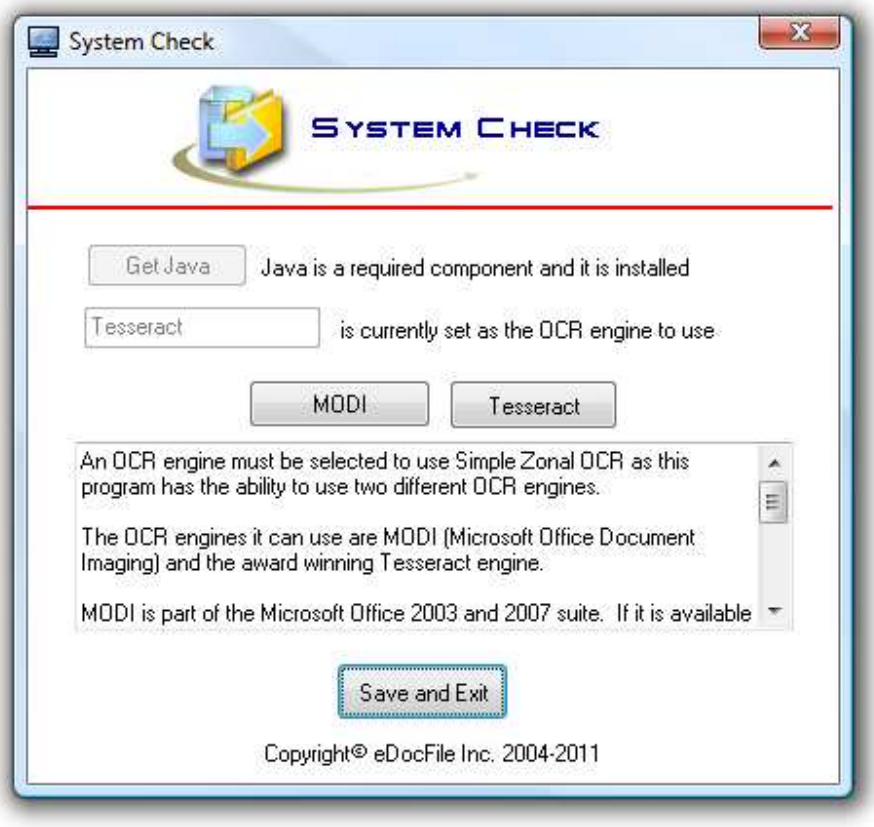
Java
Microsoft C++ 2008 run times
Either MODI (Microsoft Office Document Imaging) or Tesseract OCR engine from Google.

Please note MODI is available in Microsoft Office 2003 and in 2007 it is not part of Office 2010.
The install of Simple Zonal OCR also installed an installer for Tesseract, it is only necessary to agree to the license terms to use it.

The program install checked for Microsoft C++ run times and installed them if necessary. (unless the installer canceled it). To check for Java, MODI and Tesseract click on the "System Check" button.



The System Check Window will appear.



Java is a required component and if it is not installed the Java button will been enabled. Clicking on it will take the user to the Java website to download the program.

The window shows what OCR engine is currently being used by the program. If the "MODI" button is enabled it can be selected to be used. If it is not enabled or the user prefers the award winning Tesseract OCR engine they can click on the button, accept the license terms and install Tesseract.

Shown in the window above, Java is present as the button is greyed out and the OCR engine is set to Tesseract.

When the user exits this Window the current displayed settings will be saved.

Some notes on the two engines:

Both OCR engines do an excellent job on numbers in a standard sized font.

When the font is unusual, or very large it is better to use MODI if available.

If the documents are being virtually created (not scanned) it is better to use Tesseract engine.

If using MODI do not check "Auto Rotate" and "Auto Straighten". To see if they are checked, open MODI, click on Tools \ Options with the Options window open, click on the tab that says "OCR" and the check boxes will be displayed.

Quick Start

Step One - Prepare Documents for Scanning

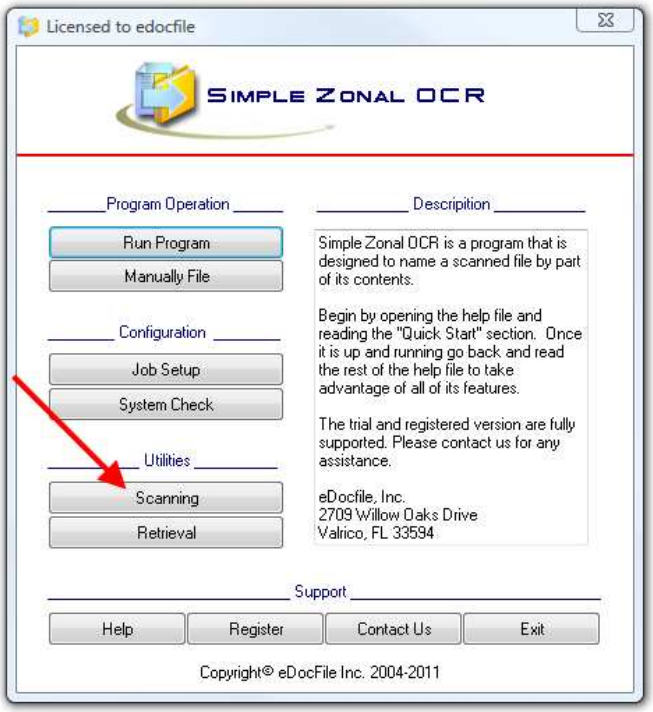
Prepare some documents to be scanned the same way they will be prepared in production. For instance if they are going to be separated with blank pages, insert blank pages. In this example multiple length invoices are going to be scanned with the use of blank pages for file separation.

Step Two - Setup Scanning

Are they going to be scanned with a network copier?

If YES, setup a destination with the images being scanned as tiff images at 300 dpi and G4 compression. G4 compression is also know as fax compression and black and white. Do not scan in color or gray scale. Once setup scan the sample file into the input folder. Go to Step Three.

If NO, Start by turning on the scanner - the software needs to find it and it has to be on to do so.

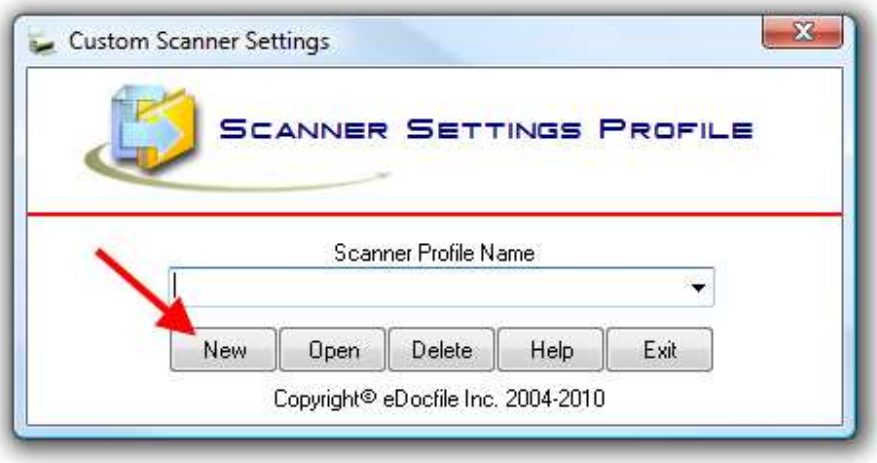


Click on "Scanning"

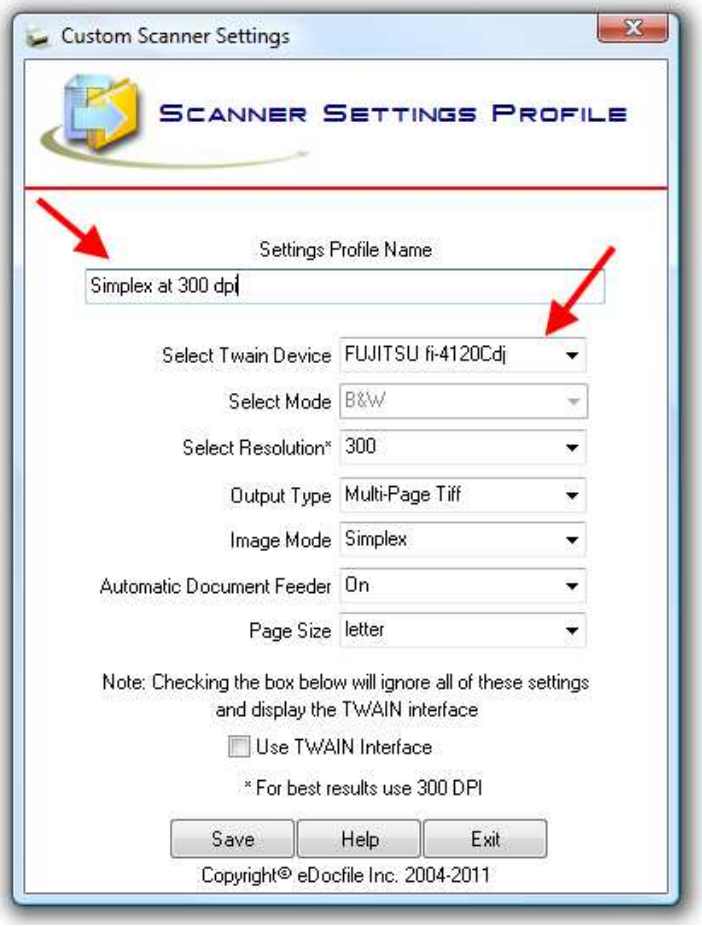
Click on "Create Scanner Profiles"

Since this is the first time the program has been run the user will be prompted to set an output folder for scanning. This Window will only appear if no path has been set. Since the scanning output path is the input for processing, to change the path in the future it is done in "Settings". Keep in mind the software works by processing files in a folder and can be used with a copier.

Enter the output path (input for processing files) and click on "Continue"


Custom Scanner Settings
SCANNER SETTINGS PROFILE
Scanner Profile Name
New
Open
Delete
Help
Exit
Copyright© eDocfile Inc. 2004-2010

Click on "New"


Custom Scanner Settings
SCANNER SETTINGS PROFILE
Settings Profile Name
Simplex at 300 dpi
Select Twain Device
FUJITSU fi-4120Cdj
Select Mode
B&W
Select Resolution*
300
Output Type
Multi-Page Tiff
Image Mode
Simplex
Automatic Document Feeder
On
Page Size
letter
Note: Checking the box below will ignore all of these settings and display the TWAIN interface
Use TWAIN Interface
* For best results use 300 DPI
Save
Help
Exit
Copyright© eDocfile Inc. 2004-2011

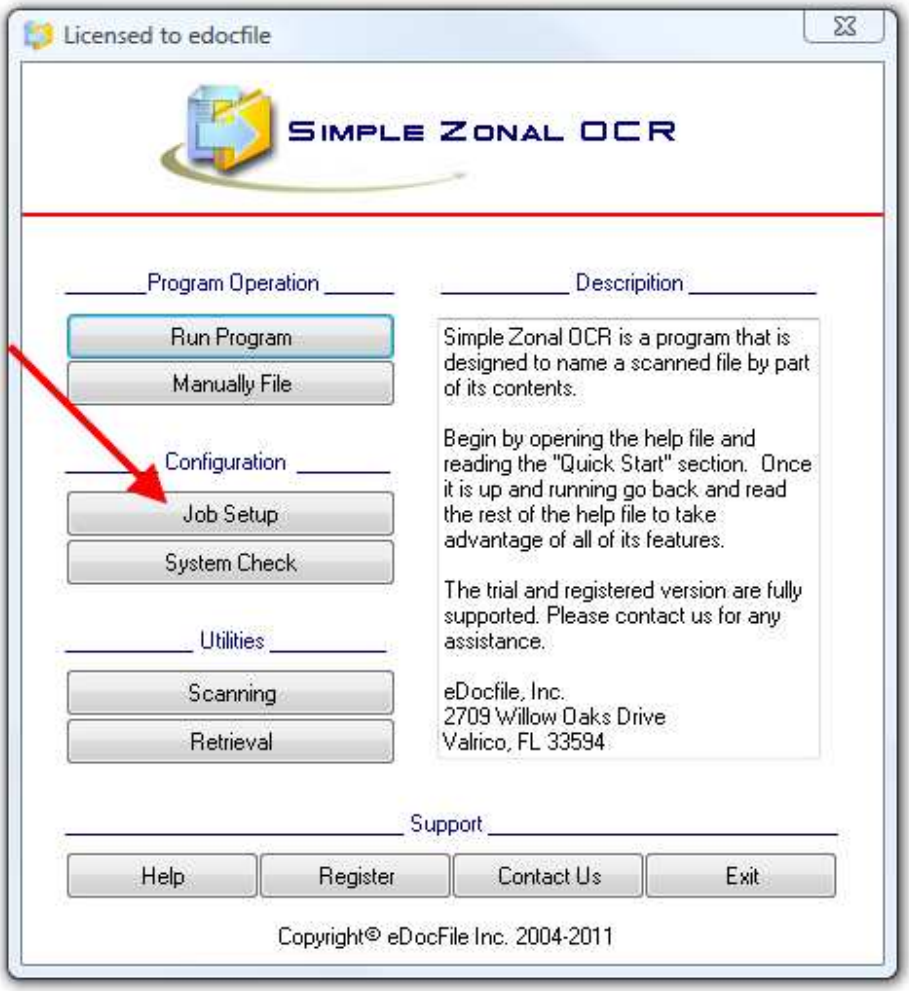
The Scanner Settings Profile Window will open. Enter a Name for this profile and select the scanner and settings to use. More on the settings are in the "Scanning the Image section". The settings show above will create a multi-page tiff image, with only the front sides of the pages "Simplex" at 300 dpi. Multi-Page Tiffs were selected as the files are going to be split with blank pages.

Click on "Save" when finished

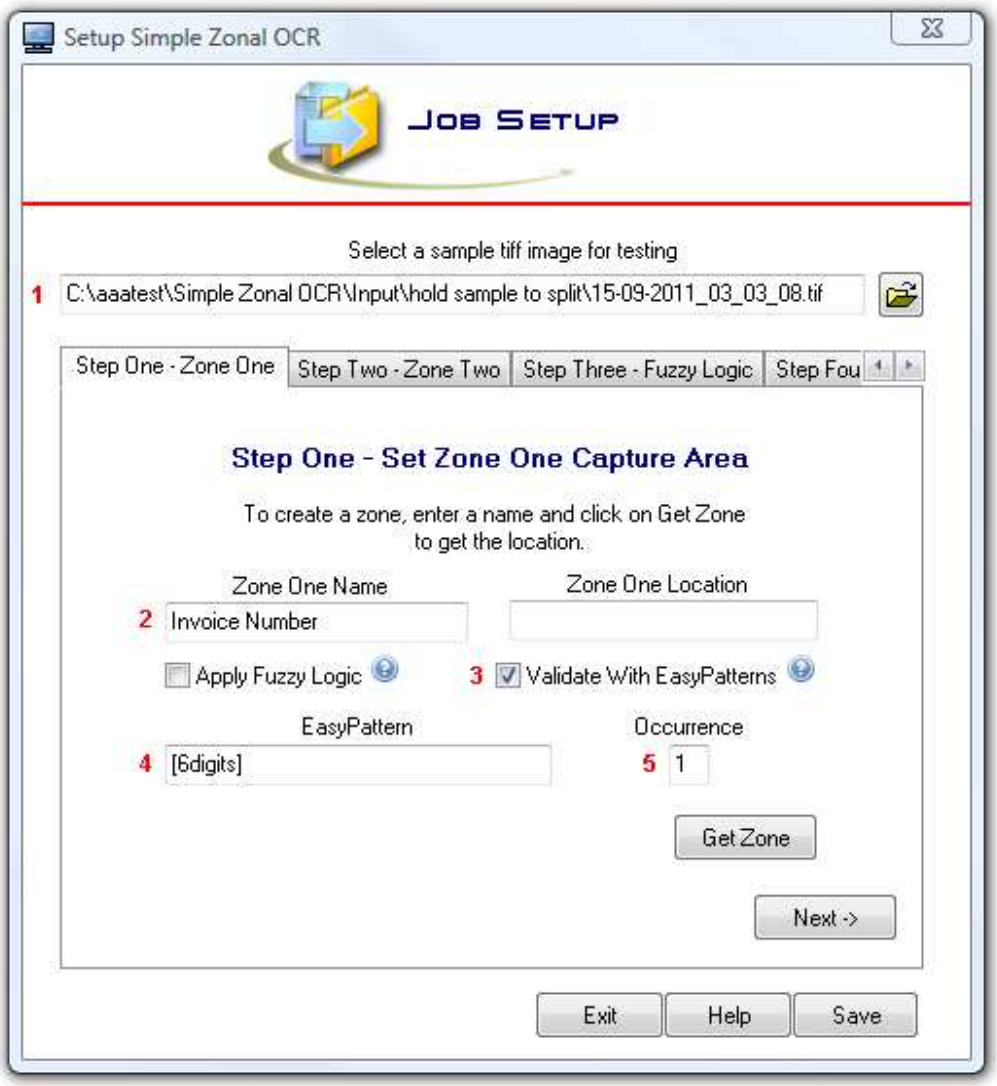
Place the paper in the scanner and click on "Scan a File"

The scanner interface will not open, paper will just go through the scanner. When the paper is finished going through and if there were no double feeds or paper jams, click on "Yes" to place the file in the input folder for processing. If it did not go through correctly, click on "No" and rescan the file.

Step Three - Job Setup



From the Main Menu click on "Job Setup"



Note: the Job Setup is broken down into 5 steps (the tabs below the sample file), since this is just a Quick Start Guide details about all the selections made in each section will be skipped.

1 - Browse to the sample file

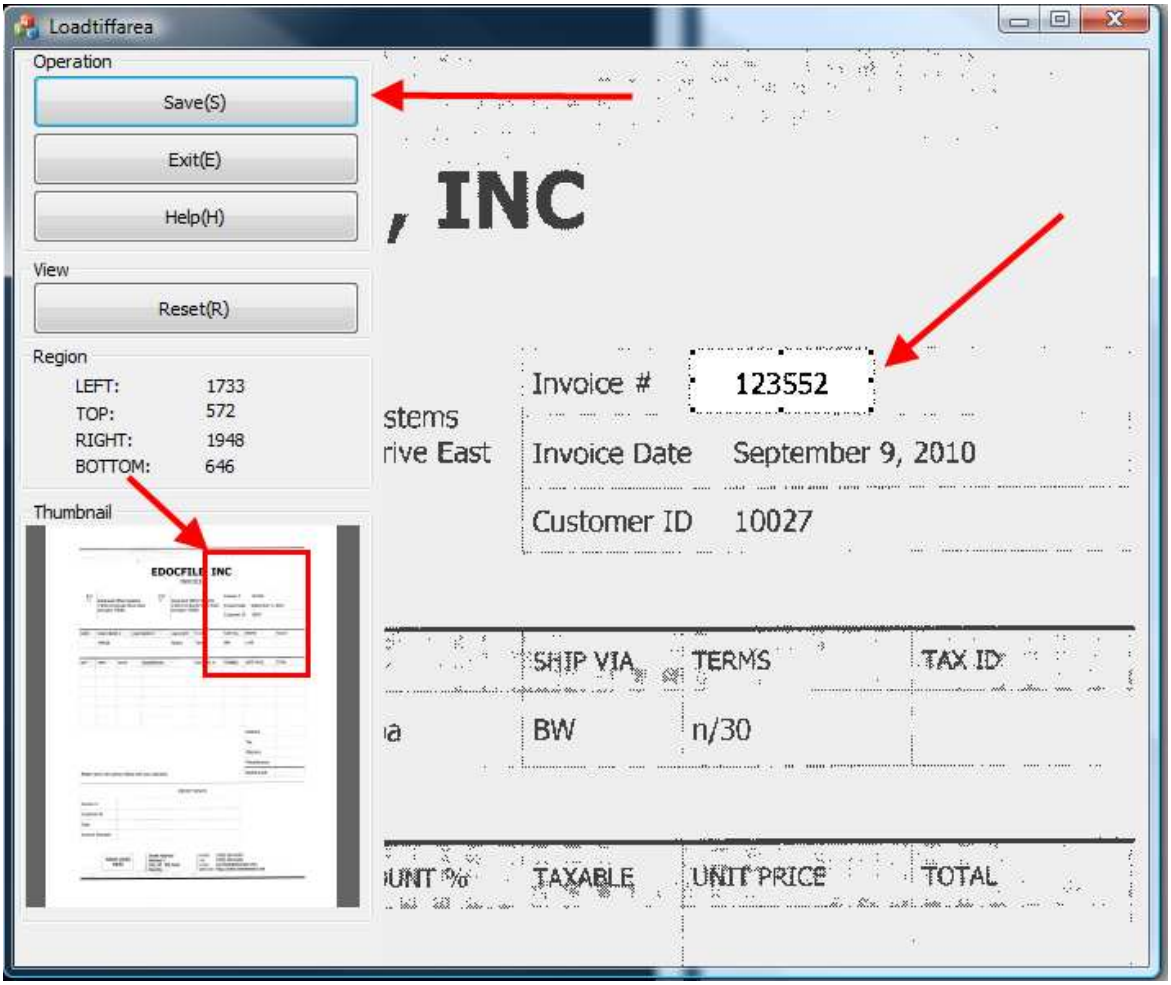
2 - Enter a name for the Zone

3 - Place a check mark in "Validate with EasyPatterns" help with EasyPatterns can be called up by clicking on the quesition mark

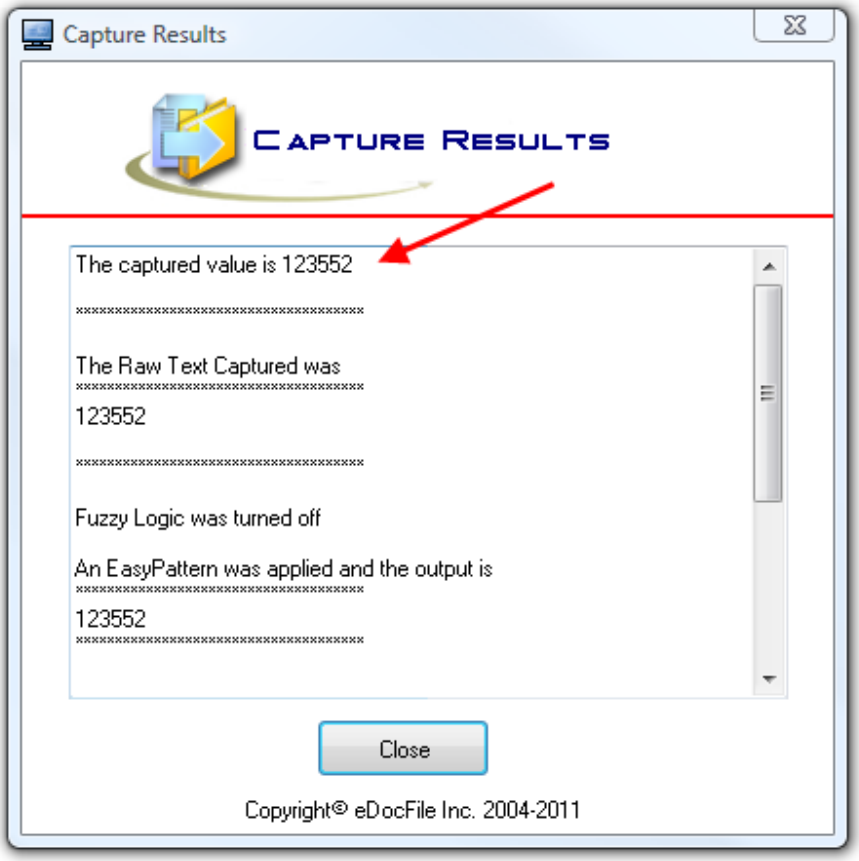
4 - Enter the EasyPattern, in this case it is a 6 digit number so "[6digits]" is the pattern.

5 - Enter the occurrence of the EasyPattern, in this case it would be the first 6 digits so 1 is entered.

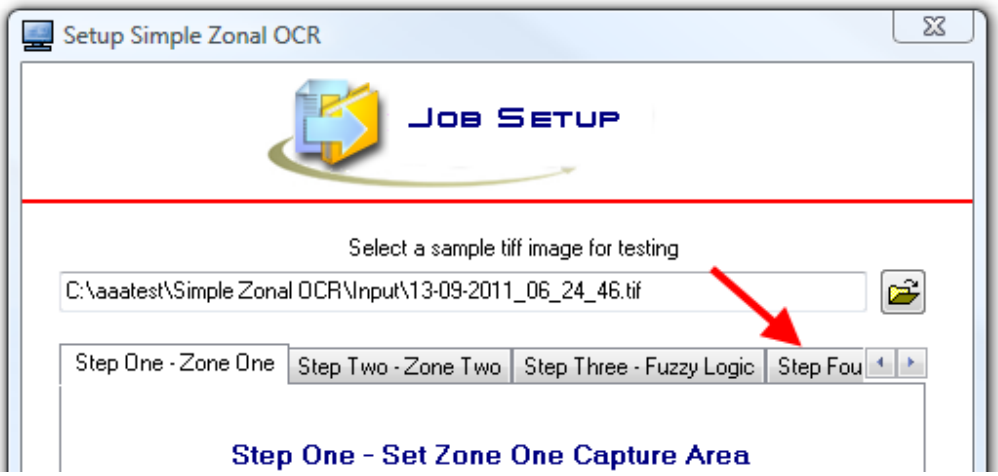
3 - Click on "Get Zone" to open the sample file and select the area to OCR



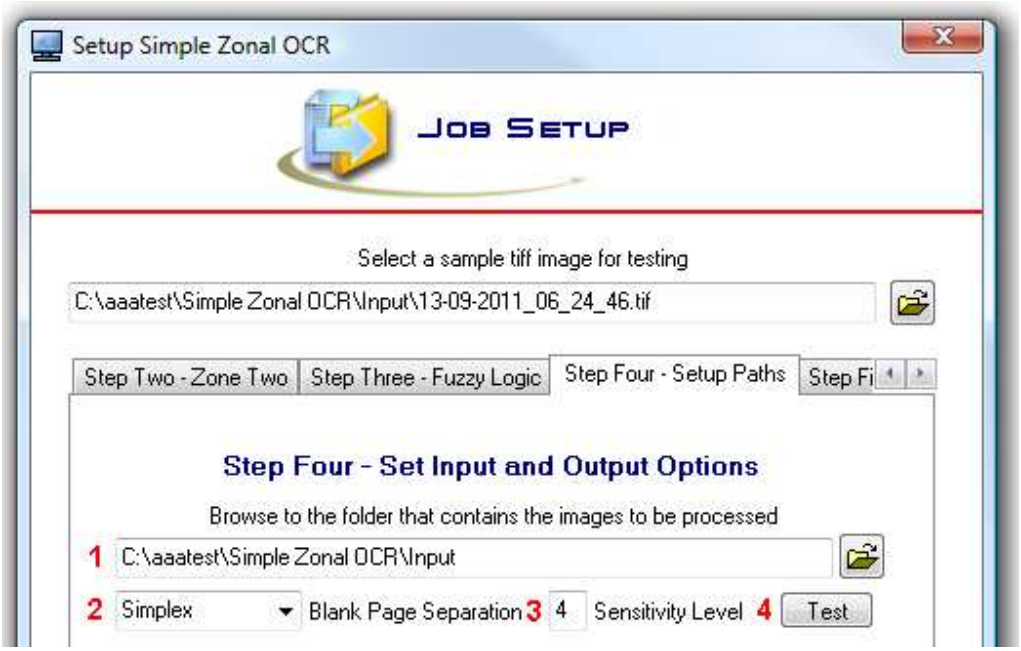
The scanned image will open in a viewer. To zoom in on a section use the mouse wheel, to move the area drag the red box on the left side, Once the area can be clearly seen, Start at the top left of the area to be captured, press the left mouse button and drag the cursor to the bottom right and release the mouse button. Once the area is highlighted click on "Save".



The captured results will be displayed check to see if they are correct. Click on "Close" to close the window. If the results are correct continue. If not, change the area by repeating the process. If the image quality is poor it maybe necessary to use "Fuzzy Logic" to correct the errors. More on Fuzzy Logic and Easy Patterns can be seen by clicking on the question marks on the window.



In this example only one zone is going to be captured and Fuzzy Logic is not being used. (it is highly recommended that they do be used to fine tune the program). Click on "Step Four" to skip Zone Two and Fuzzy Logic setup.

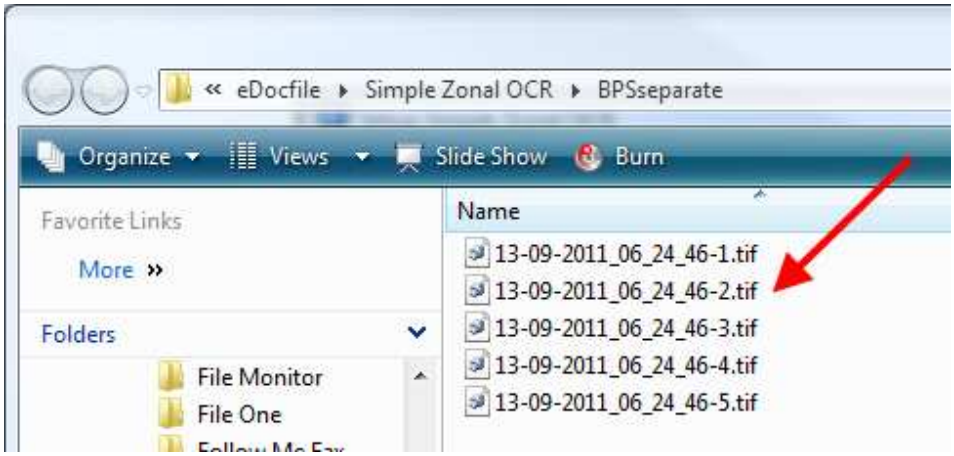


1 - Set the input folder (this will already be set if using the scanner).

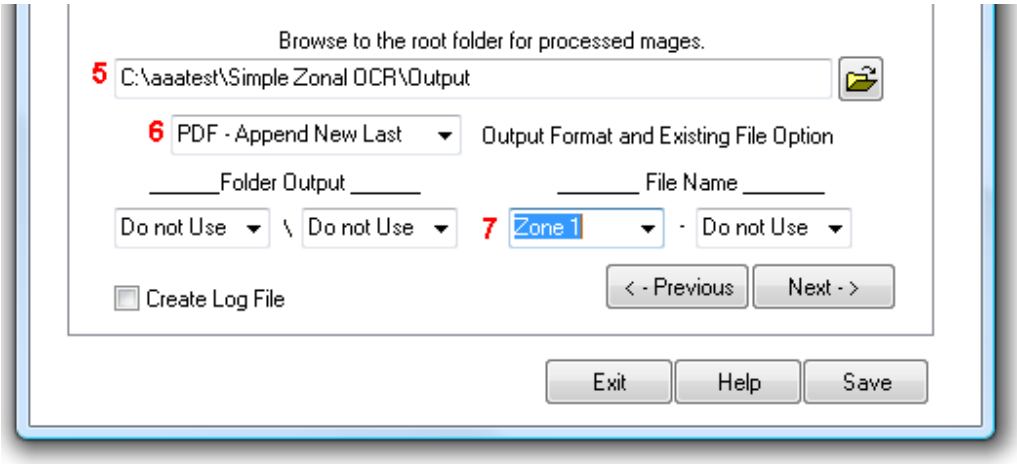
2 - If using Blank Page Separation select which type "Simplex or Duplex". Simplex splits the file each time a blank page is found; duplex splits each time 2 or more blank pages are found.

3 - Set a Sensitivity Level for splitting. The level is based on 100 percent being 1000, so setting it at four as shown here makes it split if less that 4 tenths of one percent of the pixels are black. When using blank page separation, the scanner must be clean and the option (if available) to place a border on the image must be turned off.

4 - Click on "Test" to test the separation.



In this case the sample file was correctly separated into five sections. Continue entering the rest of the settings or if they were not split correctly, chane the Sensitivity Level and try again.

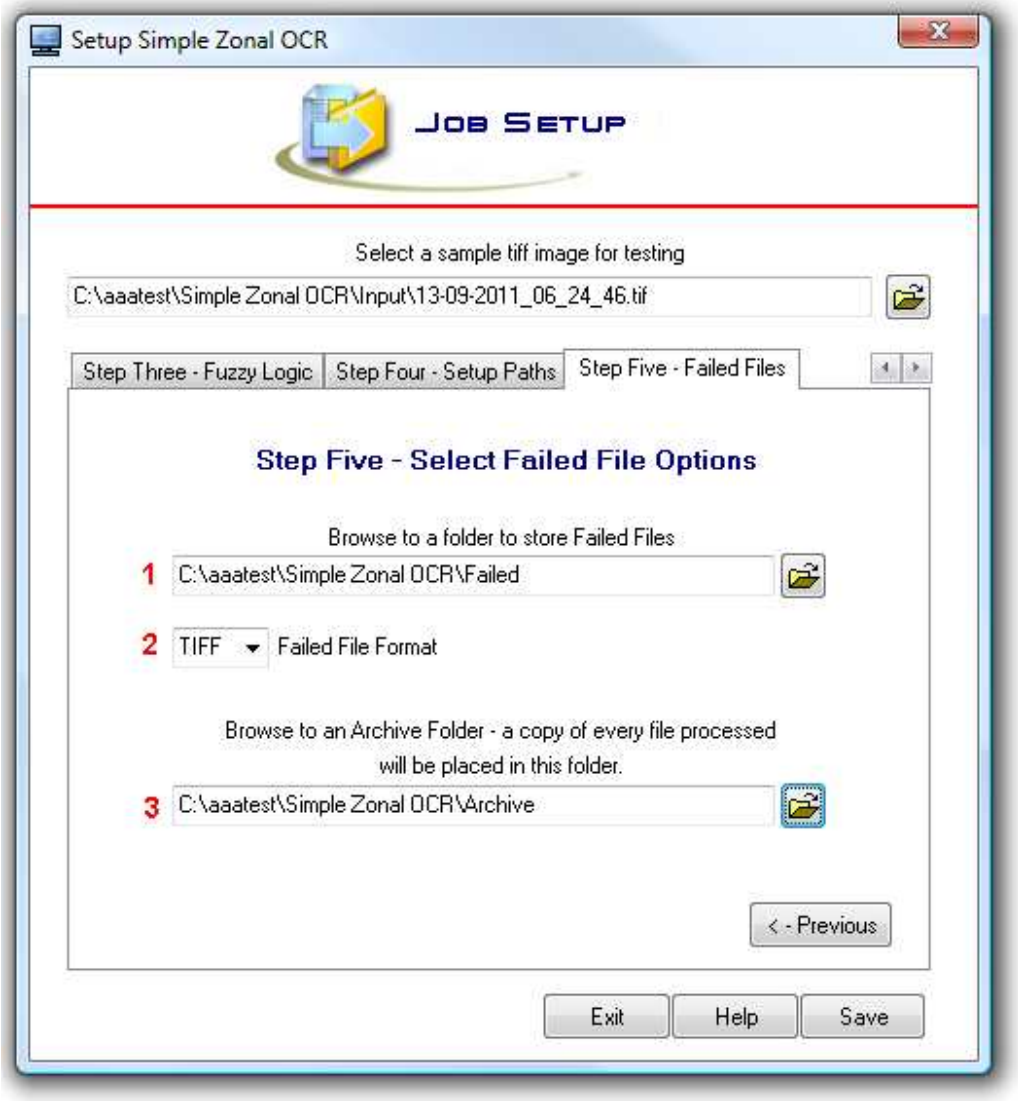


5 - Enter a folder for the processed images.

6 - Select what type of output (Tiff or PDF) and what to do if a file already exists with the same name.

7 - Select "Zone 1" to use as a file name.

When finished click on "Next" to go to the final step.



1 - Enter a folder for files that failed. (Keep in mind that this is for when the file fails to match an EasyPattern)

2 - Select a format for Failed Files. Even though PDFs can be selected, Tiffs are recommenced as they can later be checked for why they failed. Also, Manual Processing is transparent to the user as to file type being processed as a PDF and Tiff look the same in the viewer.

3 - Enter a storage location for a copy of every file that was placed in the input folder.

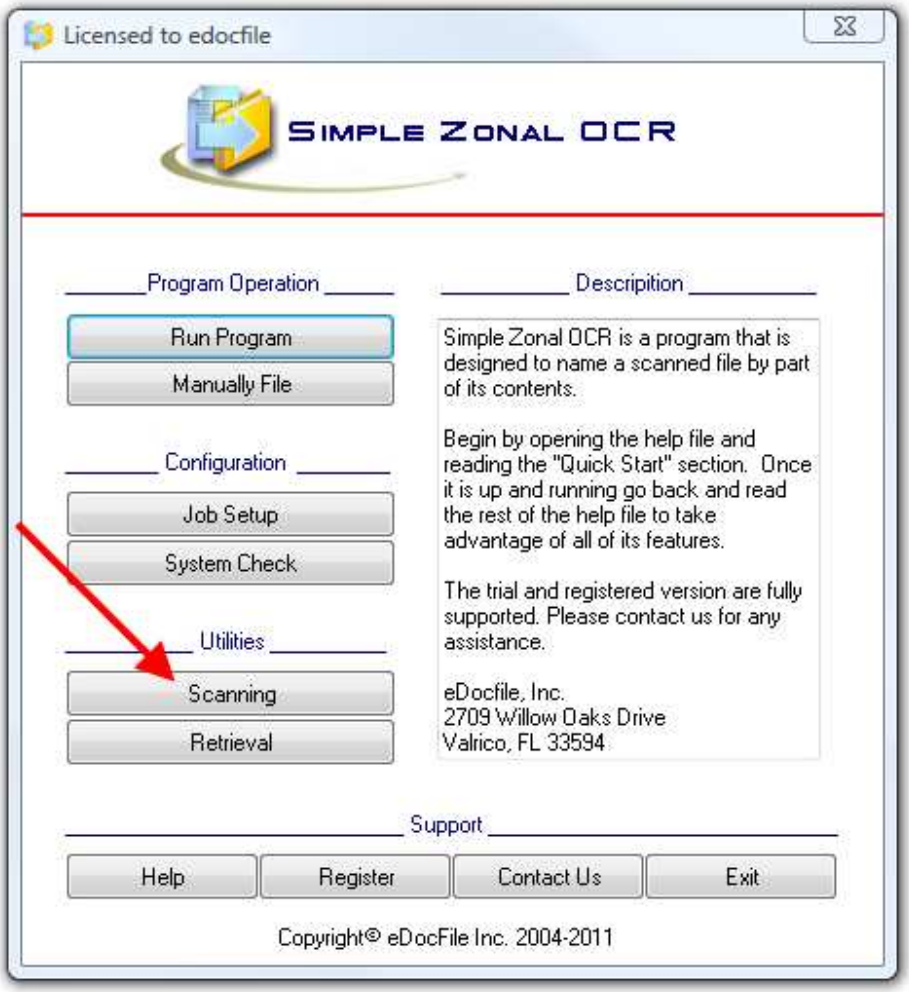
When finished click on "Save" to return to the Main Menu and the "Run Program" to begin processing.

Scanning

Are the documents going to be scanned with a network copier?

If YES, setup a destination with the images being scanned as tiff images at 300 dpi and G4 compression. G4 compression is also know as fax compression and black and white. Do not scan in color or gray scale. Once setup scan the sample file into the input folder.

If NO, Start by turning on the scanner - the software needs to find it and it has to be on to do so.

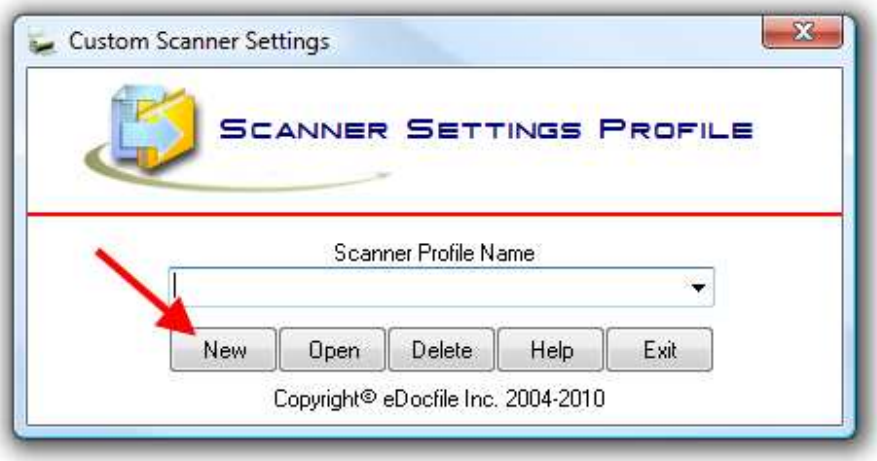


Click on "Scanning"

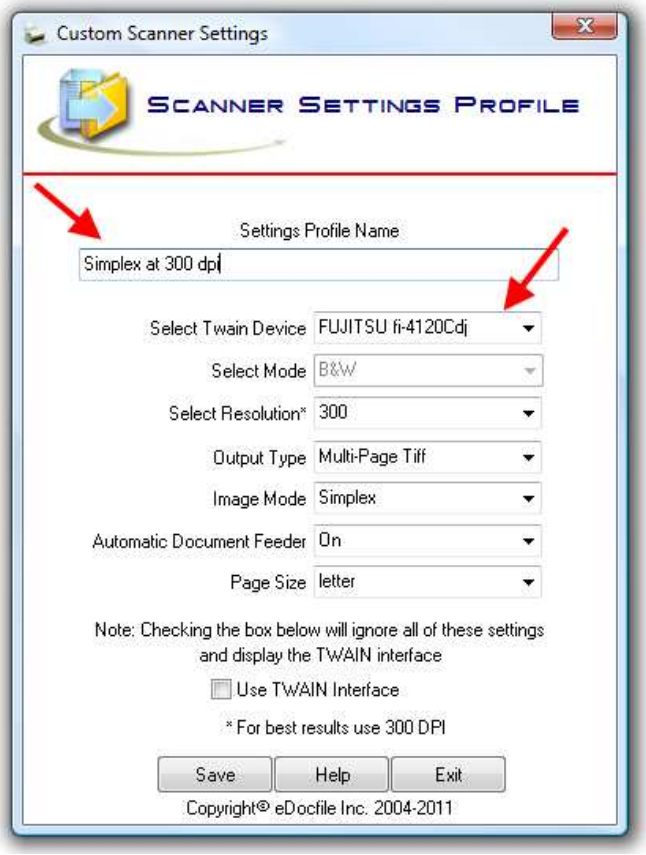
Click on "Create Scanner Profiles"

If this is the first time the program has been run the user will be prompted to set an output folder for scanning. This Window will only appear if no path has been set in the Job Settings. Since the scanning output path is the input for processing, to change the path in the future it is done in "Settings". Keep in mind the software works by processing files in a folder and can be used with a copier. Because of this the output for scanning will always be the input for processing.

Enter the output path (input for processing files) and click on "Continue"



Click on "New"



The Scanner Settings Profile Window will open. Enter a Name for this profile and select the scanner and settings to use. The name should be something relevant to the scanner settings because the software allows for more than one set of settings. For instance there maybe some single page simplex documents that come in and some multi-page documents as well. If the first page is the same the user could scan the single page ones as single page files and then use blank separator pages and scan the rest as a multi-page file. When they are processed the multi-page file will be separated and the single page ones will remain as single page files. So a description such as "Single Page Simplex Files" would be suitable. Keep in mind that if using more than one scanner setup the resolution must be the same.

The resolution should be set at 300 dpi however better results maybe had at a higher resolution if the text is smaller.

The output type can be either Single or Multi-Page Tiff files. If using blank page separators it must be set as a Multi-Page tiff image. This is the scanner output being setup not the programs output, because of that if PDF output is desired it is set in the Job Settings.

The Automatic Document feeder should be set to "On" to hide the scanner TWAIN Interface, if one is not present select No and place a check mark in "Use TWAIN Interface".

Click on "Save" when finished.

If necessary repeat the process with different settings.

Place the paper in the scanner and click on "Scan a File", if there are more than one Scan Settings Profile a drop down list will appear, just select the profile to use.

The scanner interface will not open, paper will just go through the scanner. When the paper is finished going through and if there were no double feeds or paper jams, click on "Yes" to place the file in the input folder for processing. If it did not go through correctly, click on "No" and rescan the file. It is also possible to view the file first before releasing it if it is a Multi-Page Tiff image.

Job Setup

The "Job Setup" is broken down into five steps to make setup easy.

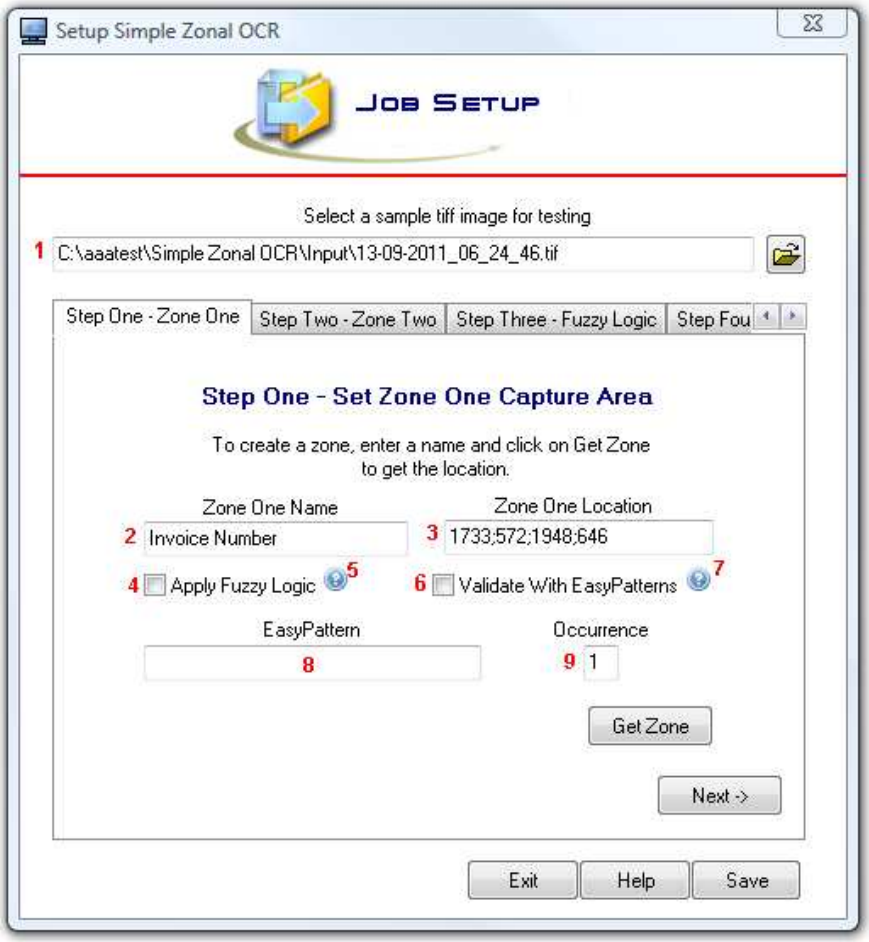
Step One and Two are where the area of the document is set to be read. On these tabs there is also quick help on EasyPatterns and Fuzzy Logic.

Step Three is where Fuzzy Logic is entered.

In Step Four and Five the file paths are entered as well as the output type and whether or not to use blank page separation.

To navigate between the steps click on the tab on top or use the "Previous" and "Next" button.

Step One - Zone One



1 - This is the sample file used for setting up the locations and for testing blank page separation.
2 - This is a name for the zone it will become a title for a data enter box for failed files.
3 - The area that is going to be captured and read. it is set by clicking on the "Get Zone" button
4 - Place a check in this box to apply Fuzzy Logic. If this is checked it will be applied before an EasyPattern
5 - Click here for quick help on Fuzzy Logic
6 - Place a check in this box to validate the data with an EasyPattern. If at all possible this should be used.
7 - Click here for quick help on EasyPatterns
8 - Enter the EasyPattern to apply
9 - Enter the occurrence of the EasyPattern.

Notes: on Zone One

"Zone One Name" and "Zone One Location" must be entered for the program to run. In other words the program will not process Zone Two if there is no Zone One.

Scan a sample file in the same manner you will use for batch processing. Ideally it will be a 300 dpi G4

Tiff image.

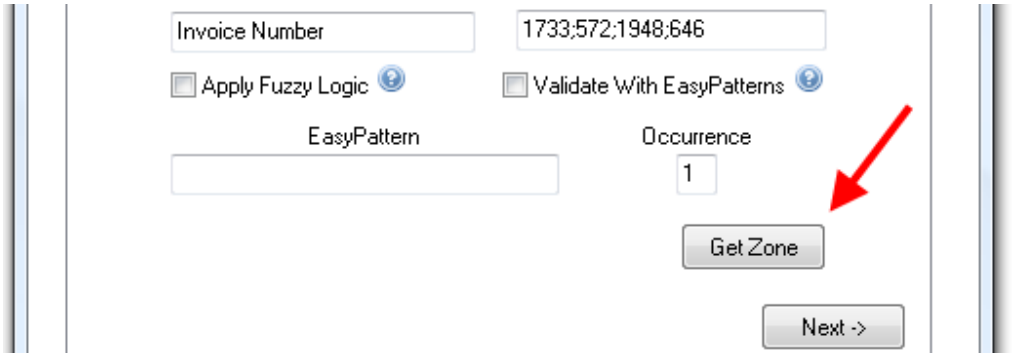
One common mistake is when using a copier to scan is the user sets the output as a PDF. This program processes Tiff images only, it can create PDF output, but it cannot process PDF input.

Fuzzy Logic should be tested before being used on anything that is not very common. For instance an OCR engine reading an "O" as a "0" would be common. What would not be is a "5" being read as a "b".

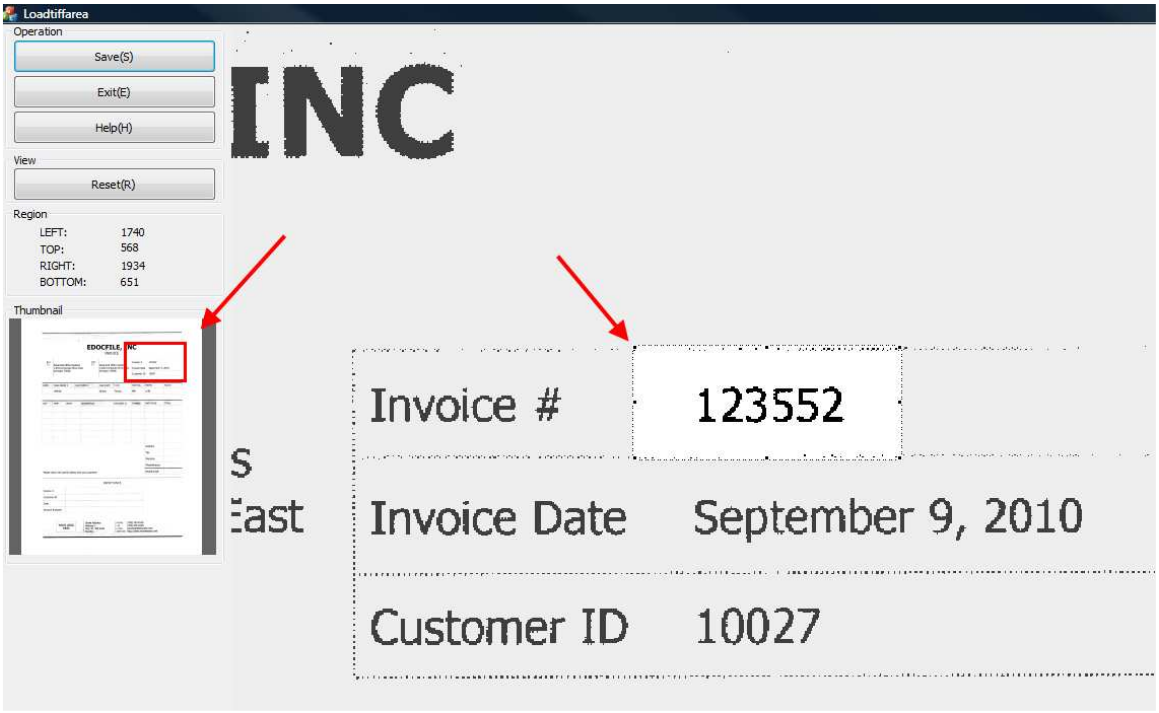
If at all possible use an EasyPattern for validation. If there is supposed to be a 6 digit number returned set it as such so that if something else is returned the file will be put in the Failed File folder for manual processing.

File separation only works with G4 tiff images. The images must be clean as the program searches for black pixels to determine if it is blank. Therefore there can be no lines down the page from a dirty scanner, borderss on the image or punch hole circles.

Getting the Area to be read



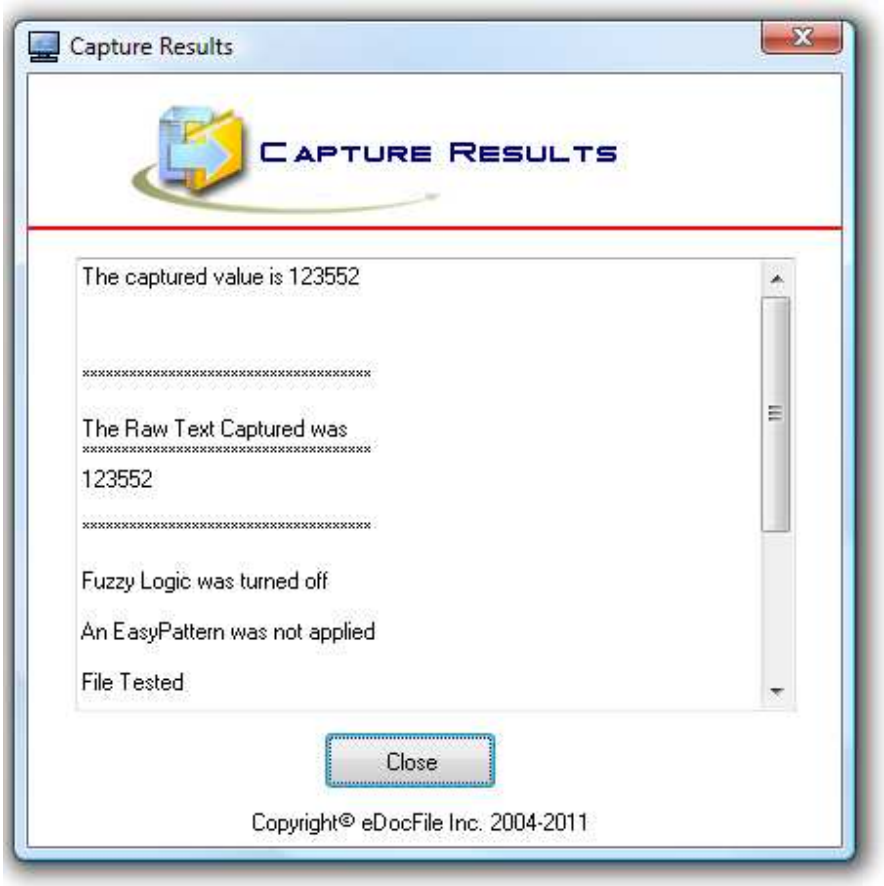
To set the area to be read, click the "Get Zone" button on either the Step One or Step Two tab.



Your sample scanned file will open in the viewer.

- To Zoom in on the image use the mouse wheel
- To navigate around the image when zoomed in, click on red box in the thumbnail view and move it.
- To highlight the area to be captured, starting at the top left hold the left mouse button down and drag the cursor to the bottom right of the area and then release the button.
- To edit an area, drag the handles on the box or just recapture the area.

When finished click on "Save" or "Exit" to exit with out saving.



The text read in the area will be displayed. When displayed it will also have the final output if Fuzzy Logic or an EasyPattern is applied.

Step Two - Zone Two

See Zone one the only difference is the quick help links

Step Three - Fuzzy Logic

Fuzzy Logic can in many cases correct failed OCR. For instance, the OCR engine could return a letter "O" when it was supposed to be a zero. This is not uncommon because of different fonts having similar characteristics.

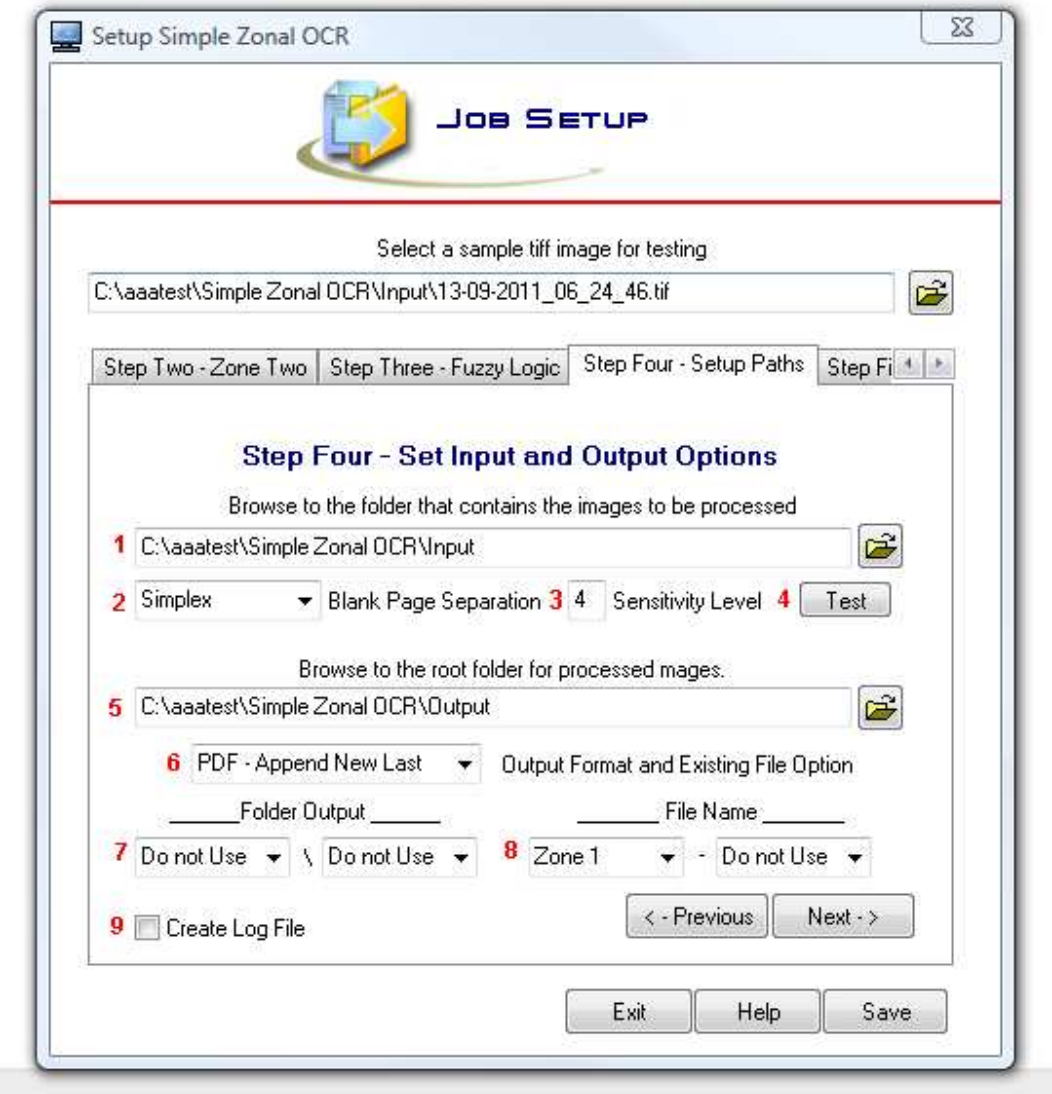
Fuzzy Logic is used primarily when numbers are being read or a specific pattern. If the zone is only supposed to contain numbers and a letter or special character is returned Fuzzy Logic will swap them out with a logical replacement.

Fuzzy Logic can be set to make assumptions such as the "O" and "o" are really zeros. Some common ones are "8" and "B", "l" and "1", "\" and "1", "/" and "1". It is not also uncommon for spaces and line feeds to be inserted when capturing an area. To setup Fuzzy logic enter what is commonly found an equals sign and what it should be.

=
o=0
O=0
I=1

To remove a carriage return line feed (part is on one, line part on another) enter %CRLF%=

Step Four - Setup Paths



1 - Enter the folder that will be monitored for tiff images

2 - If blank page detection is to be used to separate files select either Simplex or Duplex. Simplex will separate them each time a blank page is found. Duplex will separate the files when a minimum of 2 blank pages are found.

3 - Set a Sensitivity Level for splitting. The level is based on 100 percent being 1000, so setting it at four as shown here makes it split if less that 4 tenths of one percent of the pixels are black. When using blank page separation, the scanner must be clean and the option (if available) to place a border on the image must be turned off.

4 - Click on "Test" to test the separation. The sample file will be separated with the settings entered and the folder containing them wil open.

5 - Browse to the root file folder of the output path.

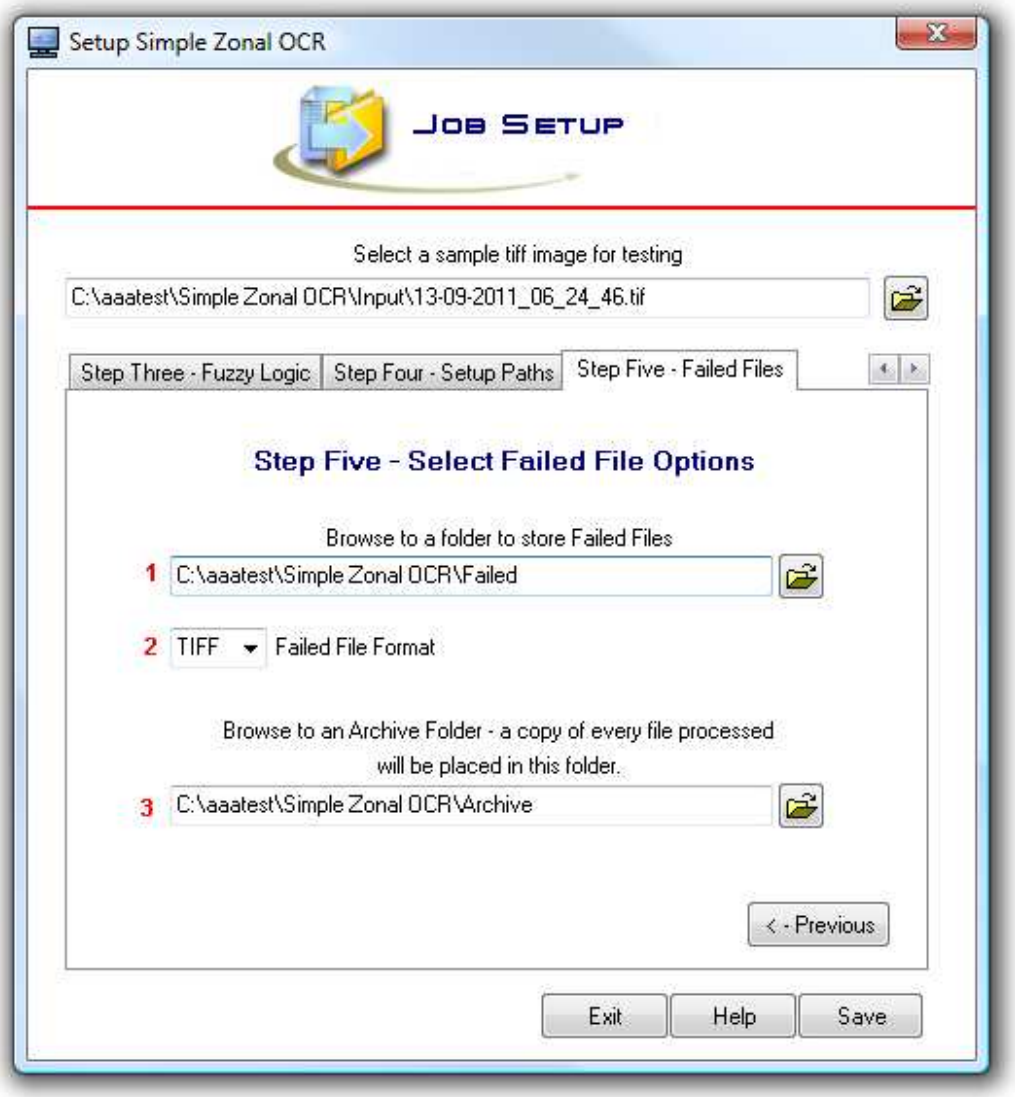
6 - Select which type of output (PDF or TIF) as well as what should be done if a file exists with the same name. The options are:

PDF - Replace
PDF - Append New First
PDF - Append New Last
PDF - Add Time Stamp
TIF - Replace
TIF - Append New First
TIF - Append New Last
TIF - Add Time Stamp

7 - If the captured text is to be used as a sub folder, select which zone should be used.

8 - Select the Zone or Zones to be used as the file name.

Step Five - Failed Files



1 - Enter a folder for files that failed. (Keep in mind that this is for later - files can only fail if they fail to match an EasyPattern)

2 - Select a format for Failed Files. Even though PDFs can be selected, Tiffs are recommenced as they can later be checked for why they failed. Also, Manual Processing is transparent to the user as to file type being processed as a PDF and Tiff look the same in the viewer.

3 - Enter a storage location for a copy of every file that was placed in the input folder.

EasyPatterns

EasyPatterns can assist in determining files that have not been correctly OCR'd. For those familiar with Regular Expressions they are very similar but much easier to understand. An EasyPattern looks for the occurrence of a certain pattern. For instance, the first six digits of a string of numbers, a letter or special character in place of a digit can all be checked for by the use of EasyPatterns

EasyPattern examples:

[6digits] - This captures exactly 6 digits

[3+digits] - This captures an entire number that has at least 3 digits

35[+4digits] - This Captures a number that begins with 35 and has 4 more digits

[('35' or '36' or '37'), 3digits] - This captures any number in the range from 35000 to 37999

[4digits]AB[2digits]VIN[2+digits] - This would return a mixture of letters and digits and the letters have to be specific. It would begin with 4 digits, have the letters AB, a two digit number, the letters VIN followed by at least 2 or more digits such as 1234AB12VIN34

[4digits][2+letters][2digits]VIN[2+digits] - This is similar to the expression above only instead of it having the letters "AB" returned it will return at least two or more letters followed by a two digit number, the letters VIN and two or more digits

More on EasyPatterns can be found here.

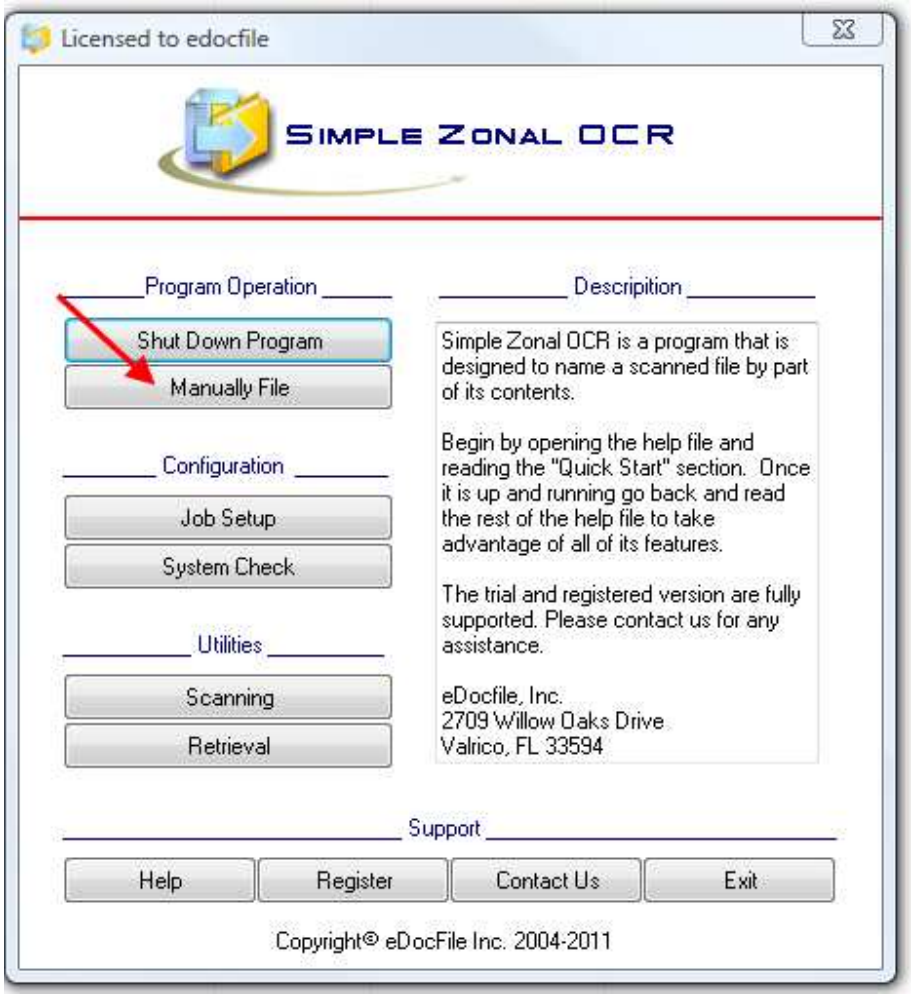
Failed Files

Not all files will be processed correctly as OCR is not 100 percent accurate. If EasyPatterns is setup to validate a filename and it fails the validation process the file will be put in the Failed File folder as either a Tiff or PDF. When first starting to use the program the output should be set as tiff. The reason for this is so that the user can browse to the failed file in the "Job Settings" and check to see why it failed, perhaps fuzzy logic could fix the error or maybe making the zone bigger could correct it.

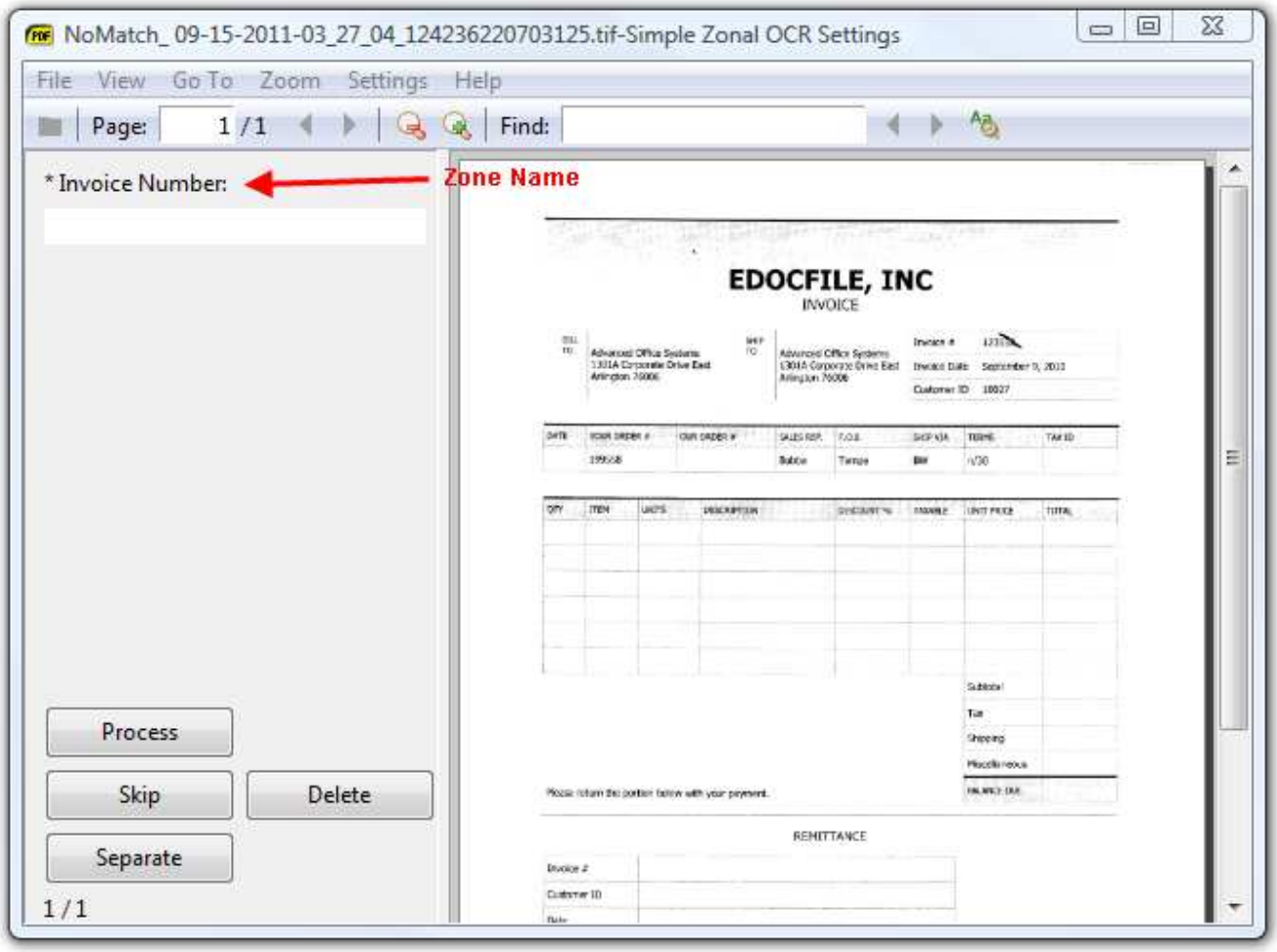
If EasyPatterns is not setup to validate a file name, all files will be placed in the standard output folder. If the text read contains characters that are not allowed in a filename the file will be renamed "non-valid-filename.(PDF or Tiff)". If the text cannot be read at all the file will be renamed "Failed_to_capture_any_text.(PDF or Tiff)".

If possible use an EasyPattern to validate the output file name.

To check for failed files click on the "Manually File" button



If there are any failed files in the failed file folder they will be opened in a viewer for manual processing. If there are no failed files in the folder the user will be informed and the utility will close. When one is in the folder this is the viewer that it will open in:



The user keys in the correct value and clicks on "Process" the next file will automatically open. The label name displayed is from the "Zone Name" in Job Settings.

Support

The Trial and Registered version are fully supported contact:

eDocfile, Inc.
2709 Willow Oaks Drive
Valrico, FL 33594
Phone 813-413-5599

support@edocfile.com