

VB-CLEAN RASTER PROCESSOR

March 20, 2001

| | |
|---|---|
| VB-Clean Introduction..... | 3 |
| Automatic Deskew Settings | 3 |
| Minimum Detect Length | 3 |
| Maximum Acceptable Skew | 3 |
| Character Protection..... | 4 |
| Automatic Rotation Settings | 4 |
| Portrait | 4 |
| Upside Down..... | 4 |
| Retain Original..... | 4 |
| Character Smoothing and Completion | 4 |
| Grow..... | 4 |
| Erode | 5 |
| Sand & Fill..... | 5 |
| Despeckle Settings | 5 |
| Despeckle Height..... | 5 |
| Despeckle Width..... | 5 |
| Diagonal Protection | 6 |
| Remove Specks with Height or Width..... | 6 |
| Dot Shading Removal..... | 6 |
| Minimum Area Width | 6 |
| Minimum Area Height | 6 |
| Horizontal Size Adjust | 6 |
| Vertical Size Adjust | 6 |
| Maximum Dot Size | 6 |
| Character Protection..... | 6 |
| Intelligent Crop Settings..... | 7 |
| Black Crop | 7 |
| Black Crop Noise Size | 7 |
| White Crop..... | 7 |
| White Noise Size | 7 |
| White Margin | 7 |
| Inverse Text Correction..... | 7 |
| Minimum Area Height | 7 |
| Minimum Area Width | 8 |
| Minimum Black on Edges | 8 |
| Invert Image Color..... | 8 |
| Invert Colors Before Processing..... | 8 |
| Invert Colors After Processing | 8 |
| Line Removal Settings | 8 |
| Minimum Line To Detect..... | 8 |
| Maximum Line Thickness | 8 |
| Maximum Line Gap..... | 9 |
| Edge Cleaning Factor | 9 |
| Character Reconstruction Width and Height | 9 |
| Manual Crop Settings | 9 |
| Top, Bottom, Left and Right | 9 |
| Crop..... | 9 |

| | |
|---------------------------------|----|
| Pad..... | 9 |
| Manual Rotation Settings | 10 |
| Turn Before..... | 10 |
| Turn After..... | 10 |
| Mirroring | 10 |
| Noise Removal Settings | 10 |
| De-Speck..... | 10 |
| Horizontal De-Speck..... | 10 |
| Vertical De-Speck..... | 11 |
| Registration Settings | 11 |
| Registration Parameters:..... | 11 |
| Resultant Left/Top Margin | 11 |
| Minimum Line Length | 11 |
| Central Focus | 12 |
| Add Only | 12 |
| Ignore Holes..... | 12 |
| Register to Line | 12 |
| Maximum Line Gap..... | 12 |
| Maximum Line Thickness..... | 12 |
| Registration Skip | 12 |
| Scaling and Resizing..... | 12 |
| Width..... | 12 |
| Height | 13 |
| Scale Factor..... | 13 |
| Sub-Image Settings | 13 |
| Top Edge..... | 13 |
| Bottom Edge | 13 |
| Left Edge | 13 |
| Right Edge | 13 |
| Pad..... | 13 |

VB-Clean Introduction

VB-Clean provides fast, accurate processing of black and white raster images. It allows groups of files to be processed in batch.

Using the various automatic features, the user can automatically rotate, crop, deskew, register and scale the images. The following is a list of the various options:

- Automatic Deskew Settings
- Automatic Rotation Settings
- Character Smoothing and Completion
- Despeckle Settings
- Dot Shading Removal
- Intelligent Crop Settings
- Inverse Text Correction
- Invert Image Color
- Line Removal Settings
- Manual Crop Settings
- Manual Rotation Settings
- Mirroring
- Noise Removal Settings
- Registration Settings
- Scaling and Resizing
- Sub-Image Settings

Automatic Deskew Settings

Skew is the angle of the text or image from horizontal. An image that is not perfectly square is considered “skewed.” The program uses text or border lines to automatically detect and remove skew from an image. The program can easily handle 10% skew, and up to 20% with a clean image. When performing Optical Character Recognition it is important to remove the skew from the image to improve the text recognition. Using the Character Protection option also improves character recognition.

Minimum Detect Length

This parameter sets the shortest line (text or graphics) the program will use to detect skew. This measurement is in pixels. This value should be at least twice the Maximum Acceptable Skew parameter. If the program reports a Failure to Detect Skew in the log file for images that should be deskewed, decrease this value and reprocess the images.

Maximum Acceptable Skew

This parameter determines the how much skew is acceptable. It is the minimum number of pixels horizontally that the program will accept for one pixel shift in the vertical direction. When the program encounters an image that shifts vertically one pixel in fewer horizontal pixels then it will Deskew the image. If there are more horizontal pixels for each pixel of vertical shift then the image is not Deskewed. Using a higher value causes the program to run the Deskew process on more of the images. Using a lower value will cause the program to accept slightly skewed images.

Use a value of at least 150 pixels for a 300 dpi image. This setting corresponds to 0.67% skew, which produces good OCR results. The setting is somewhat counter-intuitive. Lower numbers indicate greater skew.

Character Protection

This feature automatically removes rotational distortion from the De-Skewed characters, since De-Skewing an image will often distort text characters. Checking the box activates character protection.

Character Protection provides improved results and a more attractive image, but is slower than De-Skew alone. Leave it off if you want to decrease the time it takes to process the image. However, with the processing speed of today's computers, it is recommended that you leave Character Protection on.

Automatic Rotation Settings

These parameters will automatically analyze an image and determine its correct orientation. The Automatic Rotation features rely heavily on recognizing the orientation of text on the image. It may not work reliably on scans of drawings. You can use the Manual Rotation Settings to force image rotation.

Portrait

By definition, an image is in portrait orientation when the text flows horizontally across the page. When the text flows up and down the page, the image is in landscape orientation. This feature checks orientation of input images to see whether they are in portrait or landscape orientation. When the program finds an image in landscape orientation, it rotates the image 90 degrees counter-clockwise into portrait orientation. The image could be upside down after rotation. To handle upside down images, see the AutoRotate - Upside Down feature. If you have no control over image orientation, you can use AutoRotate - Portrait and AutoRotate - Upside Down simultaneously.

Upside Down

This feature checks input images to see if they are upside down in portrait orientation. When the program determines an image is upside down, it rotates the image 180 degrees to right-side up. If you have no control over image orientation, you can use AutoRotate - Portrait and AutoRotate - Upside Down simultaneously.

Retain Original

This feature forces the program to output an image in its original orientation, undoing the effects of any other AutoRotate features after completing all other processing.

Character Smoothing and Completion

The Smoothing and Completion features allow you to adjust blocks of characters (pixels, actually). Using these functions, you can bring out faint text, de-emphasize bold text, join dot matrix text into something readable by an OCR engine, and smooth out the edges of jagged text.

Grow

This feature looks at every black pixel in the original image, and adds additional pixels around it. A Growth value of 2 causes a single black pixel to become a 3 by 3 square of black pixels (adding 2 pixels, one on each side, to each direction: horizontal, vertical, and along both diagonals). This has the effect of making single pixels "grow" into small dots, and expands the borders of text characters.

Grow is very useful for making a font bolder. By growing small groups of pixels, this feature can actually "connect the dots" of dot matrix print. However, too high a value can fill in the open spaces of an 'a' or 'e' character. Lower this value if your characters start touching one another, or use Custom Grow and Erode, below, to get a finer control over the process.

By default, the program processes Growth before Erosion. This can change sparse dot-matrix print into characters resembling those from a laser printer. This can dramatically increase OCR accuracy on sparse dot-matrix print, raising recognition rates from near 0% to over 95% in some cases.

To process Erosion before Growth, use a negative value for Growth.

This process, often referred to as Dilation, is frequently used in conjunction with Erosion. When used after Erosion, it is possible to separate characters joined by thin lines (relative to the horizontal thickness of the characters) without damaging the characters.

Erode

This feature looks at groups of black pixels and removes the outer edges. This effectively makes a speck or text character “shrink” by the selected value. Setting Erosion to 2 will erode a 3 by 3 block of pixels to a single central pixel, or make a 2 by 2 block of black pixels disappear entirely.

Erosion has the effect of “peeling” off the outer layer of the character and is useful for “thinning” bold characters. Decrease this value if your characters are breaking up, or use Custom Grow and Erode. When used after Grow, it is possible to repair broken characters (provided the distance across the missing area is smaller than the distance to neighboring characters). Negative values are not allowed with this feature.

Sand & Fill

This feature smoothes the edges of characters, improving the image's appearance and reducing its compressed image file size. It fills small pits in the edges of a character, and removes small bumps on the edges.

This parameter sets the limit of “sand and fill” smoothing: increase the value to repair larger pits and bumps. The larger the type size, the larger this parameter can be. Settings between 1 and 5 work well, although you may try higher settings if needed.

Despeckle Settings

Despeckling is the process of removing small groups of pixels that are not touching anything else in the image. These specks are usually caused after scanning a faint or dirty drawing.

NOTE: Large values can cause the program to remove decimal points and periods.

Despeckle Height

The Despeckle Height parameter specifies the height of specks to remove in pixels. If Despeckle Width and Despeckle Height parameters are the same, when the Despeckle Height parameter is changed the Despeckle Width parameter will be changed also to preserve the “squareness” of the box.

Despeckle Width

The Despeckle Width parameter specifies the width of specks to remove in pixels. Use this parameter to remove oblong shaped specks which fall into a rectangle.

Diagonal Protection

In normal processing, only the four pixels above, below, and to either side of a suspected speck are evaluated. With Diagonal Protection checked, the four corners will be looked at as well. Use this feature if the Despeckle process is removing the diagonal portion of text characters or diagonal lines.

Remove Specks with Height or Width

When combined with a rectangular speck box, the program will look at the rectangle in either direction. If a speck fits with the rectangle in either a vertical or horizontal orientation, the program removes it.

Dot Shading Removal

Black and white images use dithering or dot shading to simulate shades of gray. This feature removes areas of dot shading from an image, including actual black-on-white shading. The program allows OCR reading of dot shaded text by automatically detecting zones of shading, determining the size of individual dots, and then removing them. The Dot Shading Removal feature handles a wide range of dot sizes.

Another benefit of removing dot shading from an image is that the file size will be reduced.

Minimum Area Width

This parameter sets the width (in pixels) of the smallest shaded zone to search for. Decrease this value in order to detect narrower dot shaded zones. Increase this value in order to reduce false detects or when all the shaded areas are wide.

Minimum Area Height

This parameter sets the height (in pixels) of the smallest shaded area to search for. Decrease this value in order to detect shorter dot shaded zones, or increase it to reduce false detects or if all dot shaded areas are taller than the default size.

Horizontal Size Adjust

This parameter modifies the horizontal component of the dot shading removal feature. Set to 1 if some wide dots remain; this causes the dot removal feature to remove dots 1 pixel wider than the size specified in Maximum Dot Size. Set to 2 to remove dots that are 2 pixels wider. Set to -1 or -2 if text is being degraded (leaving horizontal white streaks). This will decrease the width of the dots being sought by 1 or 2 pixels, and will also spare small text. The default of 0 works best in most cases.

Vertical Size Adjust

This parameter modifies the vertical component of the dot shading removal feature. Set to 1 if some tall dots survive the Dot-Shading Removal process; this causes the dot removal feature to remove dots 1 pixel taller than the size specified in Maximum Dot Size. Set to -1 or -2 if text is being degraded (leaving vertical white streaks). The default of 0 works best in most cases.

Maximum Dot Size

This parameter determines the maximum size of individual dots that make up the dot shading. The program will remove dots up to, and including, this size. The default value is 5 pixels.

Character Protection

These radio buttons let you select character protection with Dot Shading removal. You can protect against horizontal cuts, vertical cuts, or cuts in both directions.

This feature is more effective on horizontal and vertical segments of characters than it is on their diagonal segments. This feature may also preserve dots that touch letters.

Intelligent Crop Settings

The IntelligentCrop functions automatically crop the image to leave behind only the required data. These functions perform image analysis to determine where the edge of the data lies in relation to the edge of the image.

Black Crop

Images often have a black border around their perimeter from the background of the scanner, or other input system. This feature detects and “bleaches” the black border around those images. It will remove black borders even if they are inches thick and uneven. This feature does not remove the border; it simply turns it white.

Black Crop Noise Size

Set the amount of noise (white pixels) to ignore. Set the value larger to ignore more noise. This value is not strictly the size in pixels to ignore. The program uses advanced algorithms to make noise verses information decisions so large values (up to 100) can usually be used without damaging text.

White Crop

The White Crop feature trims the white space in the border of the image where no data appears. This feature actually trims the border to the image size. The program performs White Crop after Black Crop, which usually gives the best results.

White Noise Size

This value sets the size of noise (or specks) to ignore during white cropping. Set the value larger to ignore more noise. Unlike Black Crop Noise Size, this value is the approximate size of the speck (in pixels) to ignore.

White Margin

The value entered is the white border size (in pixels) that remains after cropping.

The program places a border around the bounding box containing the image data. You can substitute this feature for Registration, but it is not as precise.

Inverse Text Correction

Inverse Text Correction automatically detects areas of inverse (white-on-black) text and converts it into normal (black-on-white) text. It will handle multiple zones of inverse text on the page and inverse zones of various shapes, such as rectangles, circles, ovals, etc. At present, OCR engines cannot read inverse text.

These parameters can be found by clicking on the *Inversion* button.

Minimum Area Height

This parameter sets the height (in pixels) of the smallest expected band of inverse text. Decrease the setting if small inverse bands are being missed, and increase it if normal text is being treated as inverse.

Minimum Area Width

This value represents the width (in pixels) of the smallest inverse (black) zone that the program will search for. Decrease this value if small inverse bands are being missed, and increase if normal text is being treated as inverse text.

Minimum Black on Edges

This parameter sets the minimum horizontal length of the shortest run of black pixels along the left or right edges of any inverse zone. Decrease this value if the program misses inverse bands, and increase it if the program treats normal text as inverse text.

Invert Image Color

Invert Image Color will automatically switch black to white or white to black in the image. It is necessary to invert the image color if the original was an actual blueprint (white lines on a blue background) and the color was not reversed during the scanning process.

These parameters can be found by clicking on the *Inversion* button.

Invert Colors Before Processing

Setting this parameter to TRUE will reverse the color interpretation for the image after processing. For example, switching black to white or white to black.

Invert Colors After Processing

Setting this parameter to TRUE will reverse the color interpretation for the image before processing. For example, switching black to white or white to black.

Line Removal Settings

The program can automatically detect and remove horizontal and vertical lines.

Removing the lines from an image can leave gaps in characters, which can seriously degrade OCR accuracy. Use the Character Reconstruction feature to improve OCR accuracy in difficult documents, such as typed forms and those with underlined words.

Line Removal reduces OCR errors, especially in images where text and lines are placed closely together (in underlined text, for example). It is also very useful for Forms Processing and Document Imaging applications.

You can specify removal of Horizontal Lines and Vertical Lines separately with independent parameters. Other than orientation, these features are identical.

Minimum Line To Detect

The line removal system can identify and remove very short lines, such as the one across a large capital T. This value sets the minimum length of a line (in pixels) to hunt down and eradicate. Be sure to set this value larger than the height and width of your text characters.

Maximum Line Thickness

Some images may have short graphic lines you that want to remove, and large text characters you want to preserve. Since extremely large text is usually thicker than the lines on forms, you can use this feature to

keep the program from removing very large text in titles. This value sets the maximum thickness of a line for detection.

Maximum Line Gap

Lines in a scanned document often contain small gaps. Use this parameter to set the maximum gap (in pixels) that the program will consider a continuous line. Activating this feature re-attaches broken segments for removal purposes. For poor quality images, such as originals from dot-matrix printers and microfilmed documents, you may set this value as high as 50. Setting this value too high may cause the program to see small print as lines if the Maximum Line Thickness is larger than the height of lower-case print.

1 works well for most clean images. With high quality scans, you can set the Maximum Line Gap to 0.

Edge Cleaning Factor

When removing a line, the program automatically removes adjacent noise. This feature is especially useful for removing poor-quality lines. The Edge Cleaning Factor determines how far from a removed line noise will be detected.

Increase the value if line shadows remain. Decrease the value if line removal degrades adjacent text. The Character Reconstruction feature can often repair text damaged by line removal.

Character Reconstruction Width and Height

The program will automatically reconstruct intersected characters after performing line removal. This parameter sets the maximum width and height (in pixels) of characters to reconstruct.

Increase this value if text is not being adequately repaired, and decrease it if you encounter erroneous reconstruction. Text larger than 14 point may require higher settings. Setting this value to 0 disables character reconstruction.

Manual Crop Settings

Manual crop can resize an image by adding or removing a specified number of pixels from the image. This function does not alter the image resolution.

These parameters can be reached by clicking on the *Crop* button.

Top, Bottom, Left and Right

The Top, Bottom, Left and Right parameters specify the number of pixels that will be removed from or added to the edge of the image.

Crop

If Crop is selected, the pixels will be removed from the edge of the image. Image data is ignored.

Pad

If Pad is selected, the pixels will be added to the edge of the image.

Manual Rotation Settings

The manual rotate functions let you rotate an image in 90 degree increments. VB-Clean can also use the Automatic Rotation Settings to automatically orient the image.

Turn Before

This setting rotates the image before the program begins processing. This is useful when the original image is in landscape orientation or upside down, since the program (and most OCR products) works best with images in Portrait orientation, where the text flows horizontally across the image. This feature rotates the image in 90 degree increments.

Each of the radio buttons selects a specific angle of rotation. You can chose from 90 degrees, 180 degrees, and 270 degrees.

Turn After

This setting rotates the image after the program finishes processing. This is useful if you want to have the output image's orientation different from the input image's, or you want to reverse the effects of a Turn Before. It turns the image in 90 degree increments.

Mirroring

Mirror the image by flipping it either left/right or top/bottom.

Noise Removal Settings

The Noise Removal features let you remove random flecks of noise from your image, and remove Dot Shading, or 'dithered' areas of an image, without degrading text.

The program breaks the Noise Removal functions into two parts: De-Speck deals with speck removal and De-shade controls Dot Shading Removal.

De-Speck

Horizontal and vertical De-Speck are closely related features that travel along scan lines looking for specks of the user specified size or smaller.

Horizontal De-Speck

The program examines each horizontal scan line (a single pixel high), and deletes any group of black pixels whose length is less than this value. This parameter sets the maximum horizontal size for De-Speck. Increasing this value will remove larger specks. Setting this value too high can remove segments of small text. To remove large areas of dot shading, use the Dot Shading Removal feature.

Checking the Protection box causes the program to check the vertical plane to determine if the suspect pixel group is part of a character. Be aware that this is a slower process and requires more memory. For improved performance, we recommend using Isolated De-Speck instead of Horizontal De-Speck with Protection.

Without Protection active, neither Horizontal nor Vertical De-Speck looks at anything in the opposite plane. As a result, the program may cut away pieces of letters with thin vertical or horizontal segments. However, this has the advantage of having very successful and efficient "whisker" (specks touching characters) removal.

Vertical De-Speck

This parameter sets a maximum vertical size for De-Speck. The program examines each vertical scan line (one pixel wide), and deletes all groups of black pixels whose length is less than this value. Increasing this value will remove larger specks. Setting this value too high can remove small vertical sections of characters.

Check the box to invoke our Character Protection feature. When active, the program checks the horizontal plane to determine whether or not the suspect pixel group is part of a character. Be aware that this is a slower process and requires more memory. For improved performance, we recommend using Isolated De-Speck instead of Vertical De-Speck with Character Protection.

Registration Settings

This feature automatically positions text horizontally and vertically on the page, providing consistent left and top margin for documents. Registration control is helpful for Document Imaging and Forms Processing applications where accurate positioning is critical, since it makes OCR zoning of text fields in forms much more reliable.

For horizontal registration, it locates the left edge of the text and positions the left edge of the image a specified number of pixels from the left edge of text (*or data*). For vertical registration, it locates the topmost edge of the text, and positions it a fixed distance from the top edge of the image.

When using Registration, you should activate the De-Skew feature, unless you know the original is perfectly straight. You can activate and set the parameters for Horizontal and Vertical Registration separately. For most operations where you require a consistent left margin, you should activate Horizontal Registration. In many cases, you can leave Vertical Registration deactivated. For applications where precise text location is important, you should activate both Horizontal and Vertical Registration.

Registration Parameters:

With the exception of the Ignore Holes feature, Registration parameters for Horizontal and Vertical Registration are identical. The only difference is the orientation. Horizontal parameters refer to the left edge of an image, while vertical parameters deal with the top edge.

Resultant Left/Top Margin

This parameter determines how many pixels will remain as a left or top margin after processing. Setting this value to 0 places text flush with the left or top edge of the image. A negative value crops the specified number of pixels from the left or top edge of the text.

Minimum Line Length

Registration normally detects lines of text to register to. Setting this value to something other than 0 will cause the program to try to detect a line to register to.

For horizontal registration: Set this parameter to a value larger than the height of characters but less than the length of the line to detect. The program will examine the image from left to right, until it locates a vertical line segment larger than the specified value, or until it reaches one quarter the image width. It then shifts the image so that the line is the distance from the left edge of the image as specified by Resultant Left Margin.

For vertical registration: Set this property to the minimum length of a horizontal line which is expected near the top of the image. This value should be larger than the width of the text characters. The program will examine the image from top to bottom, until it locates a horizontal line segment larger than the value specified in this property, or until it reaches one quarter the image height. It then shifts the image so that the line is the distance from the top edge of the image as specified in the Top Margin property.

Central Focus

This feature causes the program to register using only the middle portion of the image border. Central Focus is useful for ignoring edge noise, letterheads and logos.

Add Only

This feature causes the program to expand the current margin (if it is smaller than the specified value) to the size specified in the Resultant Left or Top Margin parameter. If the current margin is larger than the size specified in the resultant Left or Top Margin parameter, no pixels will be removed.

Ignore Holes

Activating this feature causes the program to ignore binder holes and other left-edge noise when analyzing the margin. Leave this feature off unless you have holes or similar noise to avoid. This feature applies only to Horizontal Registration.

Register to Line

Turning on this feature allows greater control over the Register to Line feature. A value should also be specified for Minimum Line Length.

Maximum Line Gap

This value specifies the length of gaps in the line to ignore as the program searches for a line to register to.

Maximum Line Thickness

The program scans across or down your image looking for lines. In some cases the program can mistake dense text for a line. This parameter lets you select the maximum thickness to consider when registering to a horizontal or vertical line, so you can set the maximum line thickness lower than the width of the text.

Registration Skip

Use this parameter to set the number of scan lines to skip before the program starts looking for valid information for registration. The skip feature is particularly valuable for dealing with unwanted data on the edge of an image. In some cases, you may want to skip over a form line or margin note that the program “sees” as valid data. In other cases, this feature can be useful to skip over scanner artifacts that look like lines or data.

Scaling and Resizing

The program can alter the size of an image to user-defined dimensions. This feature does not change the resolution (dpi) of your image, but it will compress or expand the image data to match the new image size.

These parameters can be reached by clicking on the *Scale/Resize* button.

Width

Sets the output width of the image. The program will magnify or shrink the image to this width.

Height

Sets the output height of the image. The program will magnify or shrink the image to this height.

Scale Factor

This option allows the magnification or shrinking of the entire image just prior to saving. The program multiplies the height and width of every object in the image, and the overall image height and width by this value. You can specify a floating point value for this parameter. A negative value changes the resolution (dpi) information in the image header file to reflect the magnification. To get the magnification number, divide the length or dpi that you want to get by the length or dpi that you have.

For example: 0.8 will change a 300 dpi 8.5" x 11" image to a 300 dpi 6.8" x 8.8" image; while -0.8 will change the same image to a 240 dpi 8.5"x11" image

Sub-Image Settings

The program can isolate a specified area from an image. This extraction occurs after Deskew and Registration, but prior to any other processing. The sub-image is then processed with all other active features. The program discards the remainder of the image. This feature is useful for applications that require only part of the document. To assure accuracy and consistent output, always set the Deskew and Registration features ON when splitting out a Sub-Image.

These parameters can be reached by clicking on the *Scale/Resize* button.

Top Edge

This value defines the top edge of sub-image to extract, measured in pixels from the top of the image after performing Deskew and Registration. Use 0 to start from the very top of the image.

Bottom Edge

This value defines the bottom edge of the sub-image to extract, measured in pixels from the top of the image after Deskew and Registration. Use 0 to signify the very bottom of the image.

Left Edge

This value defines the left edge of the sub-image to extract, measured in pixels from the left edge of the image after processing Deskew and Registration. Use 0 to start from the left edge of the image.

Right Edge

This value defines the right edge of sub-image to separate, measured in pixels from the left edge of the image after processing Deskew and Registration. Use 0 to signify the image's right edge.

Pad

This value sets the width of a white border to add to all four sides of the extracted sub-image.