4 TIPS AND TECHNIQUES

The first three chapters of this user's guide focus on information that enables you to understand, install, and use the basic functions of TextBridge.

This chapter describes methods to maximize TextBridge OCR results. Specifically, this chapter covers the following topics:

- Getting the best text recognition
- Making document processing more efficient
- Saving page images
- Running TextBridge OCR from other applications

GETTING THE BEST TEXT RECOGNITION

TextBridge OCR software achieves a consistently high level of character recognition accuracy over a wide range of documents. However, there are some actions you can take to help TextBridge do the best possible job of character recognition for a particular document.

This section offers some suggestions for optimizing text recognition. It covers these topics:

- Use and maintain your scanner properly
- Adjust scanner brightness
- Adjust for colors
- Use the fax filter
- Use the word verifier
- Process multiple documents separately
- Use the Invert command in Preview

Use and maintain your scanner properly

How you use and maintain your scanner can make the difference between a successful and unsuccessful scan. Follow these tips:

- **Know your scanner.** Read and understand all documentation that came with your scanner.
- **Maintain the scanner.** Keep your scanner clean and dust-free. Keep your scanner's glass platen (flatbed) free of dirt or marks that might be captured during scanning.
- **Load the scanner correctly.** Make sure your document is not scanned at an angle. This makes character recognition more difficult.

  When using the document feeder, make sure paper guides are aligned properly for the pages you are scanning (Figure 4–1).

  If you are using the flatbed, make sure the page image is flush against the platen, and is straight. Sometimes the actual image is skewed relative to the paper it is printed on. Correct for this as much as possible before scanning.

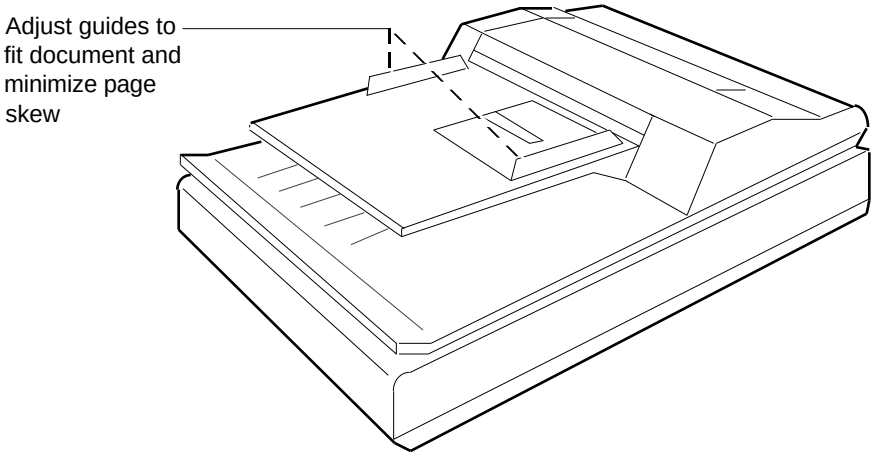

Figure 4–1. Page placement in the scanner

Adjust scanner brightness

During scanning, one of the most important settings affecting successful character recognition is scanner **brightness**. As Figure 4–2 illustrates, the original documents you scan may vary considerably.

Darkness of text, the lightness of the background, and the amount of **noise** (dirt, smeared ink, fingerprints, handwriting, and other marks) on the page can all affect character recognition accuracy.

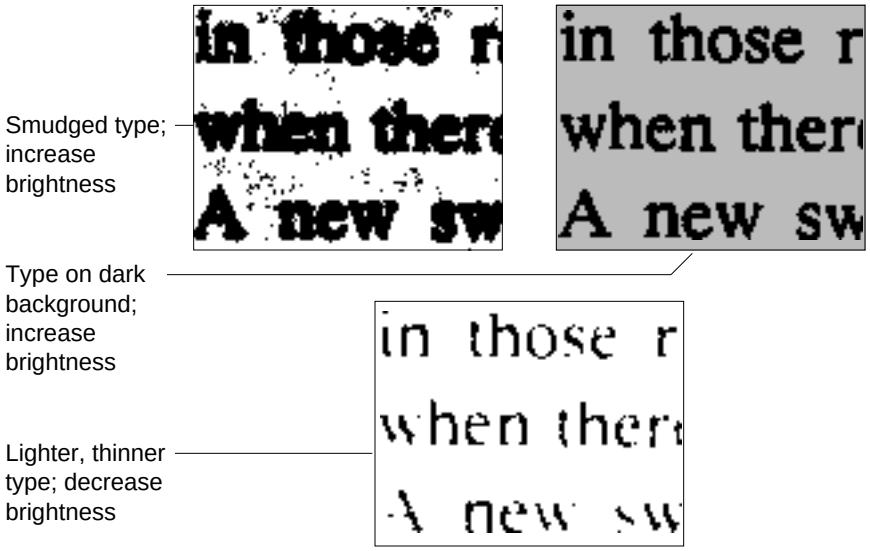

Figure 4–2. Document originals and brightness

From the Scanner Settings dialog (Figure 4–3), you can adjust Brightness to compensate for the print quality and document background. You access Scanner Settings from the Preferences dialog.

Try the Lighter brightness level if characters on your page appear too bold, are starting to fill in or are touching, or words are separated by very small spaces (as in some magazines). Recognition of documents with background noise, or with screened or colored backgrounds, can improve considerably by increasing the brightness setting.

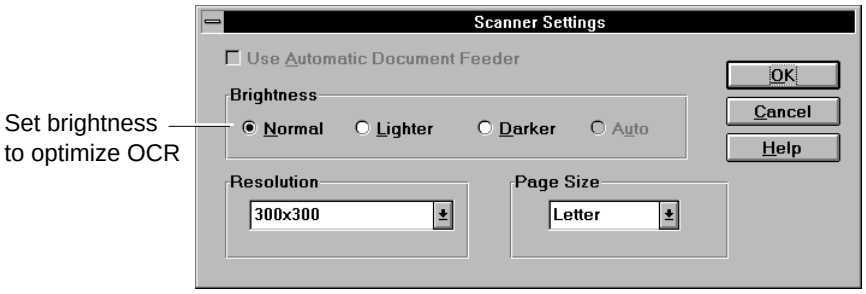

Figure 4–3. Scanner brightness settings

Try the Darker brightness level if characters on the page appear faint, broken, or very thin.

If your scanner supports the Auto brightness setting, select this to achieve the best level of brightness for each page of a document.

Another way to tell if brightness is adequate is to preview a page and zoom in to full resolution in the Preview window. This, in effect, lets you view the scanner output that the system "sees" (Figure 4–4). If the previewed image does not appear to have the proper brightness, you can adjust the setting and rescan the document.

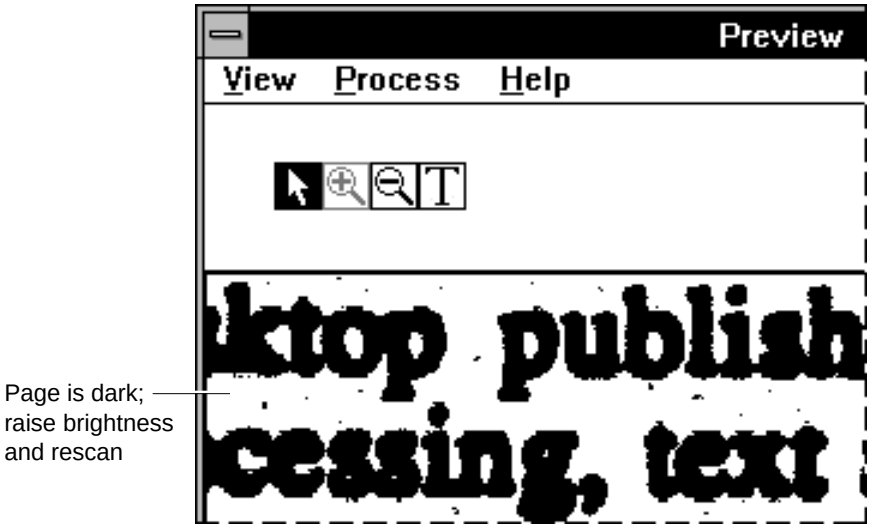

Figure 4–4. Preview display magnified

Adjust for colors

All scanners have one or more colors that they do not read. These are called **drop-out colors**. Refer to the documentation that came with your scanner to determine the drop-out color.

+ If your scanner documentation does not mention the drop-out color, examine the color of the scanner light as is moves across the flatbed. The color of the light determines the drop-out color. Many scanners have a light green scanner light, for example; thus the drop-out color would be light green.

In addition to drop-out colors, there may be other colors with which your scanner has difficulty. If the text (or image) you are scanning is colored, or is printed on a colored background, you can try adjusting the brightness setting.

If that does not work, try photocopying the page and scanning the black and white copy.

Use the fax filter

One application of TextBridge is to recognize the text in fax images. Fax images, typically, are low resolution (100-by-200, 200-by-100, or 200-by-200 dots per inch). Even so-called “fine resolution” faxes at 200-by-200 dpi are often only marginally legible (Figure 4–5).

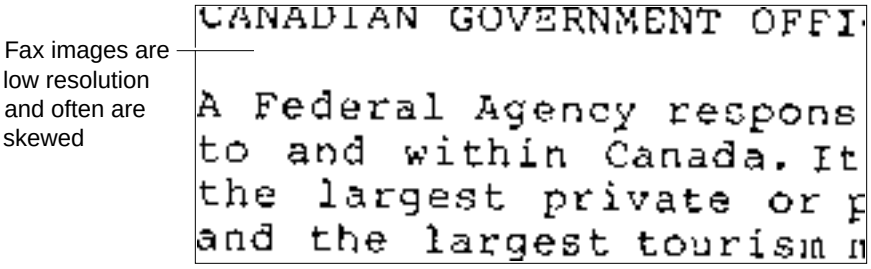

Figure 4–5. Fax image

To recognize fax images, TextBridge provides a **Fax** setting in the Preferences dialog. This Document Quality filter initiates a pre-processing step that enhances the fax image before OCR begins.

The Fax switch works on fax images stored in TIFF files **and** hard copy faxes scanned at higher resolutions (for example, 300 dpi).

Note Do **not** use the Fax filter on non-fax documents, either scanned or on-line. If you do, OCR accuracy may degrade. Also, if you notice that recognition is poor on synthesized fax images (for example, a word processor document "printed" to a fax modem), turn off the Fax filter, and try verifying part of the text during OCR.

Process multiple documents separately

TextBridge OCR software uses a variety of artificial intelligence techniques to recognize text.

With those techniques, TextBridge actually teaches itself about what it is recognizing,

Thus, TextBridge can improve OCR accuracy and speed as it scans and recognizes subsequent pages of a document.

However, you can compromise this learning capability by processing pages of different documents to the same output file.

TextBridge expects the second and successive pages of a document to use the same fonts it recognized on the first page (Figure 4–6).

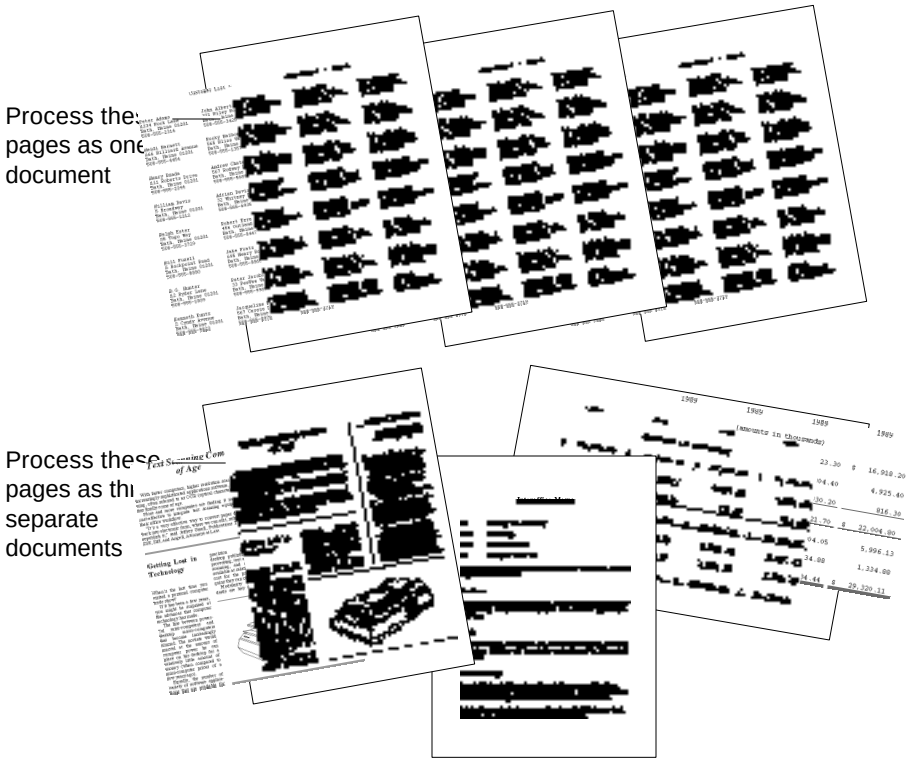

Figure 4–6. Processing multiple documents

If the second page is a totally different document, with different typefaces and point sizes, the knowledge that TextBridge gained for the first page becomes invalid.

TextBridge must begin the learning process over again for the second (and successive) pages.

If you want to scan multiple documents, and get the best recognition results, scan each document as a separate job.

Use the word verifier

If you find that TextBridge is giving less than satisfactory results on a particular document, use the word verifier to improve recognition accuracy.

By interacting with the OCR process in the Verifier window (Figure 4–7), you teach TextBridge about the characters and words in the document. This can significantly improve recognition accuracy.

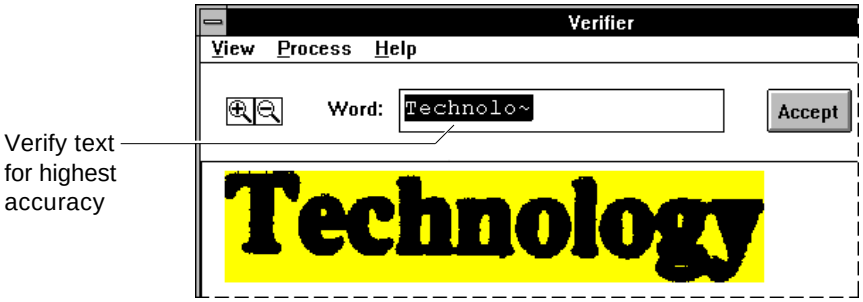

Figure 4–7. Verifier window

Each word that TextBridge is unable to recognize or is unsure about appears in the Word edit box at the top of the Verifier window. The image of the word is highlighted below for context.

With the Verifier, you can move through the recognized text and accept or correct TextBridge's recognition decisions. Your input helps TextBridge improve recognition as the job progresses.

Generally, in a multiple-page document, verify one or two pages, then end verification. TextBridge will use your input to make better recognition decisions for the rest of the document.

On small (one- or two-page) documents, to attain the highest recognition accuracy, you can verify the entire document.

For more information about using the word verifier and all its options, refer to Chapter 3.

Use the Invert command in Preview

TextBridge is capable of recognizing on-line TIFF files that originate from fax modems or other sources.

Occasionally, image data is saved so that the picture elements (**pixels**) in the resulting file are reversed: the white page background is black and the print on the page is white. This is often true with Intel FAXability files, for example.

TextBridge **cannot** recognize such files. For recognition of an on-line TIFF file, it is critical that the image contain black type on a white background.

To enable recognition of files with reverse images, TextBridge provides the Invert command in the Preview window's View menu (Figure 4–8).

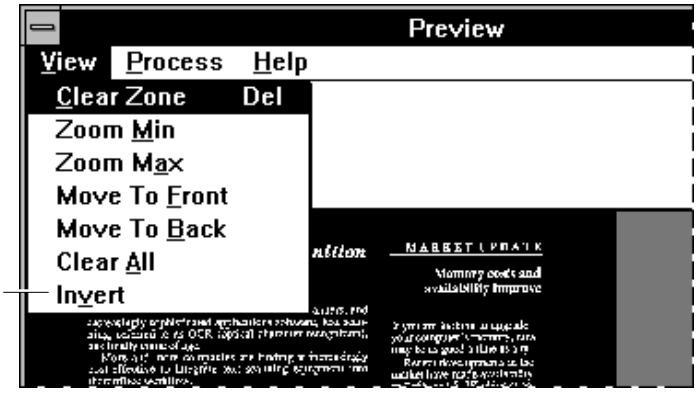

Figure 4–8. Inverting a document

If you are unsure whether an on-line file is reversed, display it in the Preview window before starting OCR. If it shows up with white type on a black background, pull down the View menu and click the Invert command. TextBridge reverses the image.

Then you can begin the OCR process. Note that inversion must be manually corrected for each TIFF file that is stored this way.

For a procedure to use Preview, refer to Chapter 3.

TIPS FOR EFFICIENT PROCESSING

When you first use TextBridge, you may find it easiest to scan a document without adjusting the default preferences. In many cases, using default preferences provides good results.

However, if you want to get the best performance from TextBridge, there are a few measures you can take before starting OCR:

- use the zone tool in Preview
- use the Ignore Photos/Halftones setting
- use auto-orientation when appropriate
- use auto-segmentation for multi-column documents

These features help to assure that the system processes only the parts of a page that are essential, and processes them correctly.

Over the course of an entire document, or many documents, using these features can translate to valuable time savings.

Zone to capture only the data you want

Some documents may display logos, graphics, running headers and footers, and other matter that you do not need to capture, and could otherwise slow down the recognition process.

With the zoning tool in the Preview window, you can identify just that portion of the page(s) that you want to capture (Figure 4–9).

Refer to Chapter 3 for information about using preview tools.

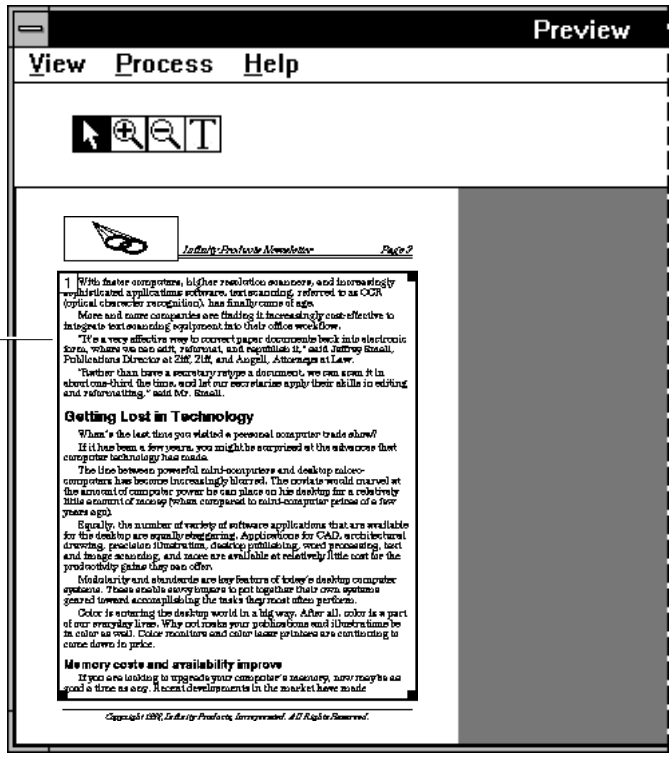

Figure 4–9. Zone in Preview

Use the Ignore Photos/Halftones option

On a printed document, a halftone photograph is made up of different-sized black dots. Ordinarily, TextBridge would spend some time trying to recognize the halftone dots as text.

Eventually, TextBridge would conclude that it was trying to recognize a halftone and would then ignore it. However, to speed up text recognition on documents that also contain halftones, you can turn on the **Ignore Photos/Halftones** setting before OCR.

With the Ignore Photos/Halftones option on, TextBridge quickly scans the page image and masks out halftones **before** beginning character recognition (Figure 4–10). Thus, actual character recognition is faster and more efficient.

Figure 4–10. Ignore Photos/Halftones filter

To use the Ignore Photos/Halftones option, select the Preferences button from the Main dialog. In the Preferences dialog, click the Ignore Photos/Halftones checkbox on.

+ Although the Ignore Photos/Halftone filtering step is relatively quick, do not specify it if your document does not contain halftones.

Use auto page orientation

TextBridge provides a tool that automatically determines the orientation of a page, rotates it in memory if necessary, then begins OCR (Figure 4–11).

Specify **Auto Page Orientation** in the Preferences dialog, which you access from the Main dialog. This feature is useful in certain circumstances, for example:

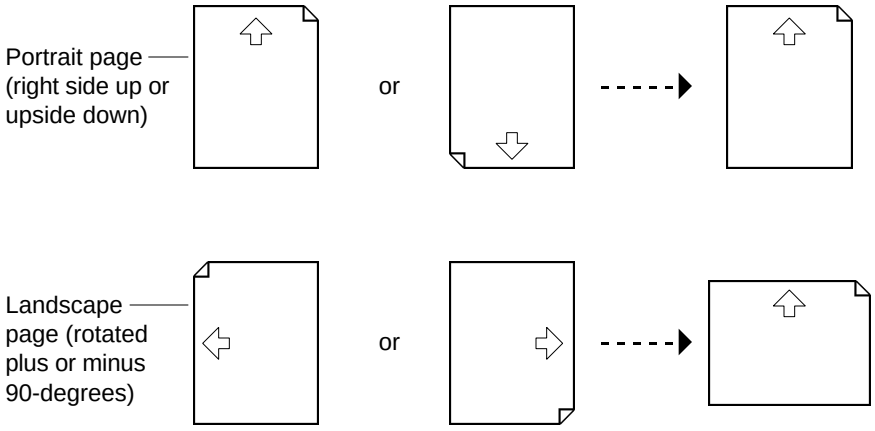

Figure 4–11. How auto page orientation works

- when you are processing documents with pages of mixed orientation
- when you are processing a TIFF file, and you do not know the orientation of the image it contains

In the first instance, you could be processing a document that has mostly portrait pages mixed with several landscape pages.

TextBridge scans each page, determines whether it is portrait, landscape (90-degrees or 270-degrees), or upside-down, and rotates it to portrait (0-degrees) before beginning OCR.

In the second instance, if the TIFF image you are about to recognize is sideways or upside-down, TextBridge will rotate it appropriately, then recognize it.

+ Auto-orientation is a processing stage that happens before recognition. Therefore, to achieve the fastest OCR, use auto-orientation **only** when the circumstances require it.

Use auto page segmentation

TextBridge provides a tool that automatically locates **regions** of text on the page, defines their order, then begins OCR. This **auto page segmentation** feature is critical for recognition of pages that have more than one column and/or unusual layouts (Figure 4–12).

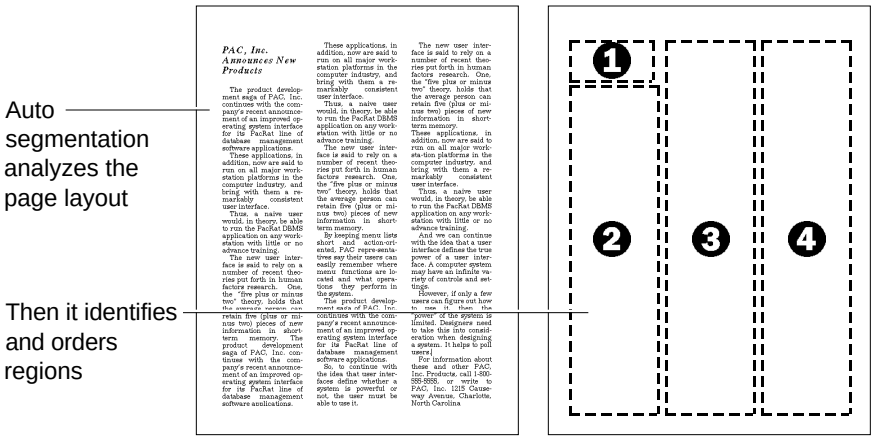

Figure 4–12. Page segmentation and region ordering

Important You **must** turn auto-segmentation **on** if you are processing pages that have more than one column of text. Otherwise, TextBridge can output regions of recognized text in the wrong order. Do not use auto-segmentation on single-column documents.

Auto-segmentation is a pre-processing stage that occurs before OCR begins.

Note that you can create a zone in Preview and still use auto-segmentation. Auto-segmentation will work inside the zone only. So, for example, if you draw a zone around two columns of a three-column layout, auto-segmentation will detect and order the two columns in preparation for the OCR process.

SAVING PAGE IMAGES

One of the checkbox options on the main dialog is **Save Page Images**. This option, available when you set Input From Scanner, enables you to save a binary (black and white) image of each page scanned during a TextBridge OCR session.

Note TextBridge saves page images as TIFF files with CCITT Group 3 compression. Group 3 is a compression standard specified by the CCITT (Consultative Committee of International Telephone and Telegraph), an international standards organization.

After you click GO!, and the first page is scanned, TextBridge displays the Save Page Images As dialog (Figure 4–13).

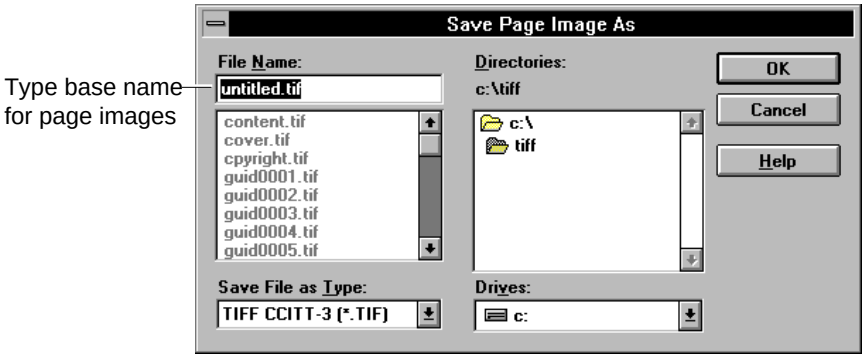

Figure 4–13. Save Page Image As dialog

The Save Page Image As dialog is very similar to a Windows standard Save As dialog, in that it allows you to specify a file name, file type, output directory and disk drive.

The file name is the **base name** on which the page image file names are built. It is also the **document name** that appears by default in the TextBridge Save As dialog when OCR is completed. The default file name is `untitled` with the `.tif` extension.

For example, suppose you specified the name “guide” in the Save Page Image As dialog. Page image files would be stored with a name in the format:

```
guidnnnn.TIF
```

where *nnnn* is the document page number with leading zeroes (for example, 0001, 0002, and so on).

Page image files are named sequentially within the directory. If a file of the same name (for example, “`guid0001.tif`”) already exists, TextBridge will start with the next number in sequence.

Also, at the end of the job, the document name (for example, “`guide`”) automatically appears in the File Name box in the normal Save As dialog.

In the Save Page Image As dialog, the initial working directory is the directory from which you launch TextBridge:

```
C:\TXBRIDGE\BIN
```

However, you can specify any other disk drive and directory in which to store the page images. This becomes the new **working directory** for the job, and, like the document name, will also be in place in the Save As dialog when OCR is completed.

For page image file format, the Save File As Type menu provides only one selection, TIFF CCITT-3 Intel. Page images are saved exactly as scanned in binary (black and white) format.

Note that if you click Cancel in the Save Page Image As dialog, the dialog closes and TextBridge terminates the job. The Main dialog remains ready for you to start again. (For example, if you do not want to save page images, you can click the Save Page Image checkbox off and re-start the job.)

RUNNING TEXTBRIDGE FROM OTHER APPLICATIONS

TextBridge OCR for Windows is actually a suite of applications that enables you to run OCR from within virtually any other Windows application.

In addition to the main utility, which runs as a standalone program, and has the widest feature set, TextBridge OCR is provided in two other forms:

- **TextBridge Application Server**, a program that acts as a menu item from inside virtually any registered Windows text application (word processor, desktop publishing program, spreadsheet, database application, and so on).
- **TextBridge OCR Printer**, a capability that enables you to send an image in any format to a version of TextBridge OCR that works like a conventional print driver.

This section provides information about using TextBridge OCR in these forms.

Note TextBridge also supports a DDE interface. Interested developers and system integrators should call Xerox Imaging Systems Customer Support for details.

Use the TextBridge Application Server

The TextBridge Application Server (TAS) is a Windows program that can be "attached" to, and thus run from within, other Windows text applications.

Once attached, TAS appears in the host application's File menu as the **TextBridge OCR** command. When you select TextBridge OCR, the TextBridge main dialog appears as if it were a dialog of the host application. From here, you can set up and initiate OCR exactly as you would with the standard TextBridge program.

Starting TAS and registering applications

TAS is installed during the TextBridge installation process described in Chapter 2 of this manual.

At the end of the installation process, the TextBridge `setup` program creates a TextBridge OCR program group that includes the TextBridge Application Server program item (Figure 4–14).

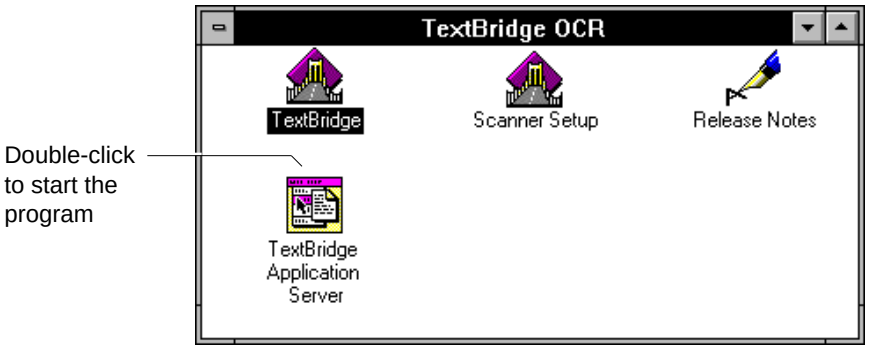

Figure 4–14. TAS program item

Before you can run TAS from within an application, you must start the program, and you may have to register the application, as well:

1. **Double-click the TAS program item in the TextBridge OCR program group.**

   The program starts and appears as a minimized icon on your Windows desktop.

2. **Double-click the icon on the Windows desktop.**

   The TAS registration dialog appears (Figure 4–15).

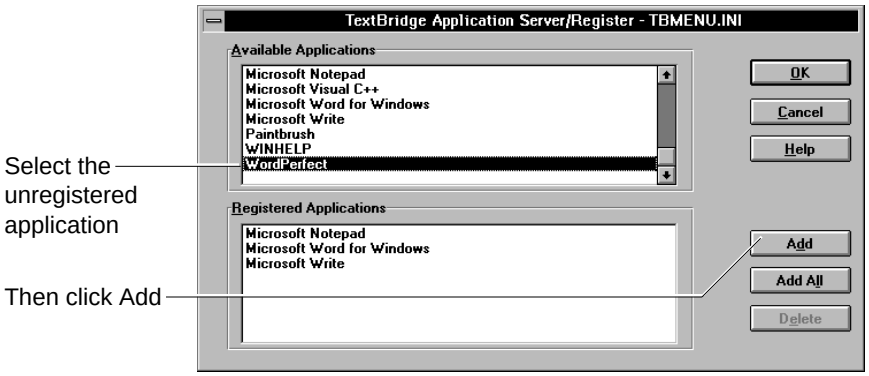

Figure 4–15. TAS registration dialog

3. Register the application, if necessary.

- At the top of the registration dialog, highlight the application that you want to register.
- Click the Add button to add the application to the Registered Applications list at the bottom of the dialog.
- When you are done registering your application(s), click OK.

You can now go on to use TextBridge OCR from within your registered application.

Running TAS from within your application

In the File menu of any active registered application, the TextBridge OCR command appears as the last command directly above the Exit command.

Note For TAS to work, the host application must have a File menu, and in the File menu, an Exit command. A majority of Windows applications use this standard.

As an example, Figure 4–16 shows the TextBridge OCR command in the WordPerfect® File menu.

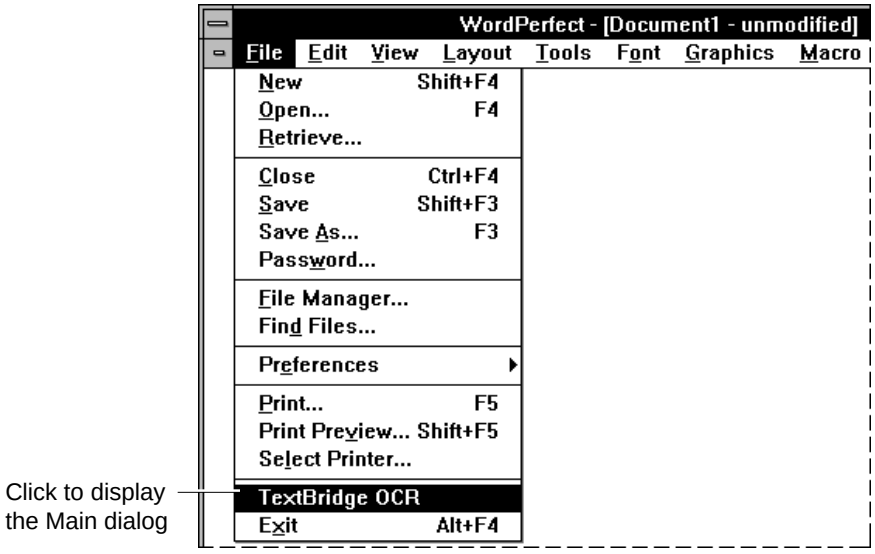

Figure 4–16. TextBridge OCR command

To run TAS, and import recognized text directly into the host application's open document, use the following procedure.

1. **Start the TextBridge Application Server.**

   Double-click the program item in the TextBridge OCR program group (refer to Figure 4–14).

   + To have TAS start automatically whenever you start Windows, place it in the StartUp program group.

2. **Make sure the host application is registered.**

   Refer to the procedure in the previous section, "Starting TAS and registering applications."

3. **Start the host application.**

   With the host application, open a new or existing document into which you want to import recognized text.

4. **Pull down the host application's File menu and select the TextBridge OCR command.**

   Status messages appear:

   ```
   Connecting to TextBridge Services...

   Connection with TextBridge Services
   established.
   ```

   In a few moments, the TextBridge main dialog appears.

5. **Set up and initiate OCR from the main dialog.**

   With the main dialog displayed, you can choose either File or Scanner as the input source, specify Preferences, and proceed exactly as if you were using TextBridge as a standalone application.

   + The Preview, Verify, and Save Page Images capabilities are **not** available in the TAS version of TextBridge. If you require these capabilities, run TextBridge as a standalone application, and save recognized text in your word processing or other text format.

   Refer to Chapter 3 for step-by-step procedures for using TextBridge; refer to earlier sections of this chapter for usage tips and techniques.

   When OCR is complete, TAS closes, and recognized text appears at the cursor position in your application's open document ready for editing.

   + TAS uses the Windows clipboard to cut and paste recognized text to your application either as formatted **RTF** (Rich Text Format) or as plain **ASCII** text. If your application supports RTF pasted from the clipboard, then RTF is used. If not, recognized text is pasted as plain text, and the formatting (bold, italic, and so on) is lost.

Use the TextBridge OCR Printer

In its other forms, TextBridge OCR for Windows can recognize image files only if they are stored in TIFF format. Certain applications, particularly some facsimile (fax) programs, store fax page images only in PCX, DCX, or some other proprietary format.

To run OCR on non-TIFF images, you can use the **TextBridge OCR printer**. The OCR Printer is designed to appear in Windows applications as just another target printer. It enables you to "print" an image from a Windows application and produce a recognized and formatted text file as the result.

A typical use of the OCR Printer is to OCR a page image directly from a fax or imaging application. The OCR printer is similar to the model of many fax programs that use a similar feature to send faxes. That is, the fax image is "printed" to the fax modem and sent to another fax modem or fax machine.

The OCR printer offers the added advantage of being able to recognize virtually any image format (DCX, PCX, Corel, TIFF, and so on). Virtually any Windows program designed to handle images can make use of the OCR printer.

To prepare the OCR printer for use, refer to the following subsection, "Adding the OCR printer." Then refer to "Using the OCR printer in your imaging application" for instructions to download an image and produce a text file.

Adding the OCR printer

The OCR Printer program files are installed along with the main TextBridge application.

However, as with actual printer drivers, you must add the OCR Printer to the list of printers available to Windows applications on your PC.

This procedure assumes that you have already installed TextBridge as described in Chapter 2.

1. **From the Windows Program Manager, open the Main program group and double-click the Control Panel icon:**

   This opens the Control Panel window with icons for various parts of the system.

2. **Double-click the Printers icon in the Control Panel window.**

   This opens the Windows Printers dialog box (Figure 4–17).

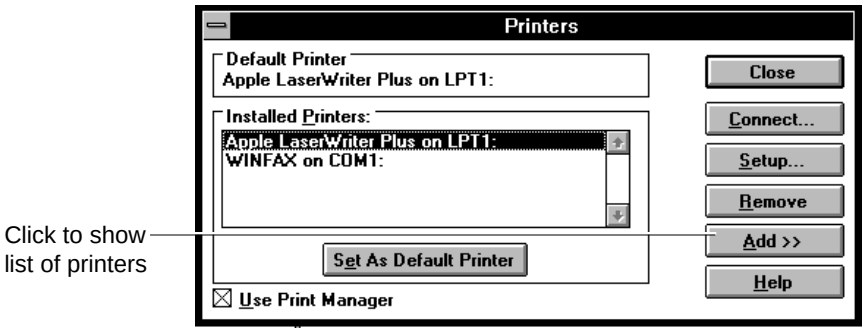

Figure 4–17. Printers dialog

3. **Click the Add button in the Printers dialog.**

   The dialog box expands to show the List of Printers that can be added.

4. **Highlight the following item in the list, then click the Install button.**

   ```
   Install Unlisted or Updated Printer
   ```

   This action displays an Install Drivers dialog box in which you are instructed to specify the drive and directory location of the printer driver.

5. **In the Install Drivers dialog box, enter the TextBridge BIN directory pathname:**

   ```
   c:\txbridge\bin
   ```

6. **Click OK (or press Enter).**

   This displays the Add Updated or Unlisted Printer dialog (Figure 4–18).

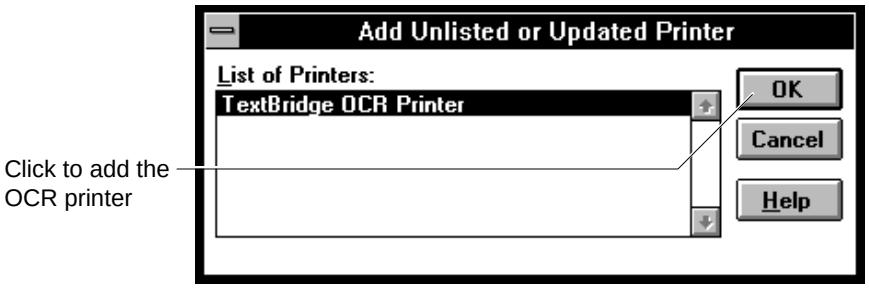

Figure 4–18. Add Updated or Unlisted Printer dialog

7. **Select the OCR Printer entry and click OK.**

   The Add Updated or Unlisted Printer dialog closes, leaving open the Printers dialog.

8. **Click Close in the Printers dialog to end the Add process.**

   You can now go on to use the TextBridge OCR Printer as described in the next subsection.

Using the OCR printer in your imaging application

You can use the OCR Printer with any Windows application that can open and view an image file.

For example, WinFax Pro (from Delrina Technology Inc.) provides an Image Viewer to view, manipulate, and print fax images. Using the Print command in the File menu of WinFax Image Viewer, you could specify and use the OCR Printer to perform character recognition on the fax image.

The OCR Printer enables you to access TextBridge Preferences before beginning recognition. After recognition is complete, you can specify the output text file name, location, and format in the standard TextBridge Save As dialog.

To use the OCR Printer in your application, follow these steps:

1. **Open the imaging application, and display the image to be recognized.**

   + The image must be binary (black and white) and within the accepted range of resolutions supported by TextBridge. TextBridge can recognize images of 100-by-200, 200-by-100, 200-by-200, 300-by-300, and 400-by-400 dots per inch.

2. **In your application's Print Setup dialog, specify the TextBridge OCR Printer as the destination printer.**

   You should have by now already added the OCR Printer in the Windows Control Panel Printer program, as described in the previous subsection, “Adding the OCR print driver.”

3. **Optionally, define TextBridge preferences.**

   - Click the Options or Setup button in your imaging application's Print Setup dialog. This displays a secondary Setup dialog.

   - Click the Preferences button in the Setup dialog to display the TextBridge Preferences dialog.

   - Specify standard or fax document quality, page orientation, auto page segmentation, and so on.

   - When you are done, click OK in the Preferences dialog and move up out of the other dialogs, as well.

     + In the Print Setup dialog of your application, if there is an option to use the actual printer resolution, turn it **on**.

4. **Start the OCR process on the displayed image.**

   - From your application, pull down the File menu and select Print.

   - Click OK in the Print dialog. Processing messages now appear as OCR proceeds:

     ```
     Processing...

     Acquiring Image...

     Recognizing text...
     ```

     When recognition is complete, the TextBridge Save As dialog appears (Figure 4–19).

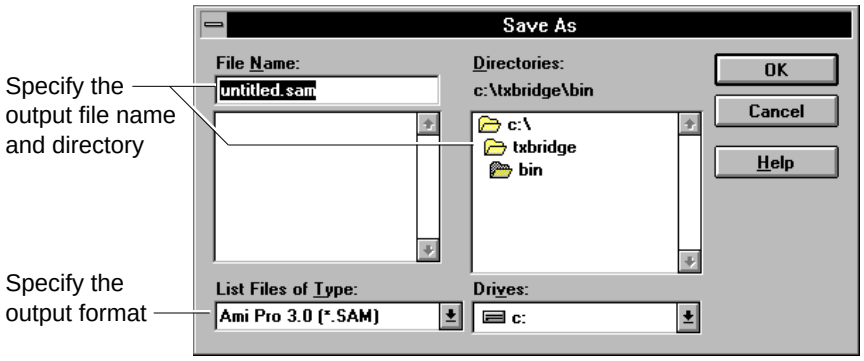

Figure 4–19. Save As dialog

5. **Specify the output file name, format, drive and directory destination, then click OK.**

   The recognized text file is converted to the specified format and written to your hard disk.