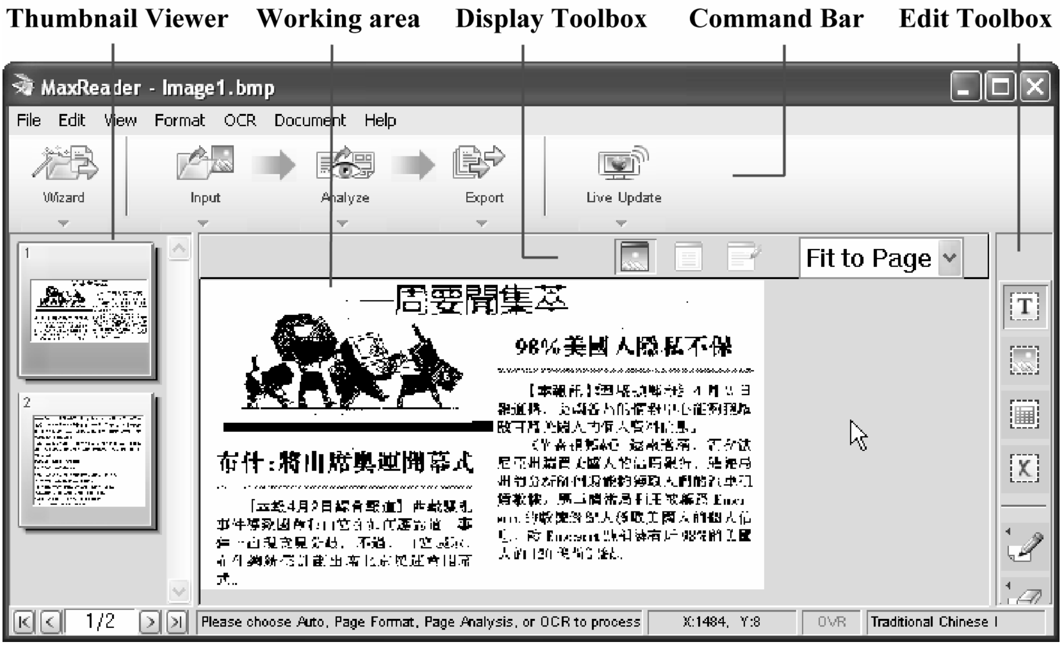

Please choose Auto, Page Format, Page Analysis, or OCR to process

Figure 4.1 MaxReader 5 Main window

To use image files as the input source, click **File**, click **Open Image File**, and then select a file or files from there. MaxReader 5 supports various image file formats, including TIF, PCX, BMP and JPG. These are more than sufficient for our needs. Even though MaxReader 5 supports the recognition of text from color images, I recommend using color images only when the colors are simple and the contrast of text to the background is high. Otherwise, it is better to manually convert the color images into black and white (B/W) before recognition (refer to the description in Chapter 3.7.3 for details).

MaxReader can take the mixing of multiple images, from the scanner and/or preexisting files. It adds all the pages to a project and initially arranges them according to the order of opening. After a page is opened or scanned from a device, it appears in the Working area and also in the Thumbnail Viewer. The Thumbnail Viewer shows a thumbnail view of all the pages in the project and you can navigate to different pages from there.

When you click the thumbnail icon of a page, the page will be selected and brought to the Working area. Use the dropdown box in the **Display Toolbox** to adjust the zoom setting for viewing the image. You can rearrange the sequence of pages in the project by dragging a thumbnail to another place. Select **Image Information** from the **Document** menu to get basic information regarding width, length, resolution, and file size of the selected image. MaxReader also comes with basic processing functions that can be used to fix up images. These functions include rotation, touching, cropping, and inverting the color of the image. MaxReader cannot recognize white (light) text on black (dark) background. For images of this nature, you must reverse the color using the inverting function prior to the recognition. Note that the inverting function provided here only allows you to invert the whole page. You must perform extra processing in a situation where an image contains both white text on black and black text on white. Cut out the region with regular text into the Windows Clipboard, do the inverting to the rest of the page, and then paste the region back. The other way is to use the GIMP 2 image editor software discussed in Chapter 3.7.

4.1.3 Settings

After all the pages are opened or scanned, and properly arranged in the project, the next thing is to check the settings and make adjustments if needed. A critical setting is the selection of the Chinese character sets. Click **Format** from the main menu, select **Set Character Set**, and then make the selection of traditional or simplified character set from there. The OCR will not work properly if this selection in incorrect. There are two choices for traditional Chinese settings. The **Traditional Chinese I** is suitable for common everyday documents, while the **Traditional Chinese II** is to be used with documents composed of ancient words or characters that appear in old poetry or literature, etc. You cannot mix pages with text from different character sets. If there are multiple documents within one project, all the text must come from the same character set. The setting of character set stays in effect until it is changed.

Select **OCR** on the main menu and then click **Recognition Preferences** to open the **Recognition Preferences** dialog window. This dialog allows you to adjust source document related properties. Certain types of block layout are interpreted by the program as multiple text tracts. The **Force to Single Column** in the **Page Layout Property** section is used to force them to be recognized as single column. My recommendation is not to use it, as this usually causes problems when you view results and perform proofreading. The **Form** section allows you to select whether the document contains forms or tables. The **Text Alignment** section allows you to select the printing direction of the text. The selections in these two sections should be self-explanatory. In the **Data Type** section, select all the text types, including Chinese, alphabetic and numeric characters, unless you are certain that some of the types will never appear in your source document. Leave the other three sections on **Auto** to let the software determine their properties for you. Click the **Set as Default** button to apply these settings. These default settings will stay in effect until you change them. If you only want to change some settings temporarily, select **Format** from the main menu and then select the **Page Format** to open the **Page Format Setting** dialog. From there you can make the same adjustments described above, but any changes made only apply to the next recognition.

MaxReader 5 can recognize many different Chinese fonts, including Song (宋), Hei (黑), Kai (楷), FongSong (仿宋), Yuan (圆), and even Li (隶书). Texts of different fonts can be mixed on a page, and the program will automatically recognize all the fonts that it supports. There is no setting or configuration to be adjusted for this.

4.1.4 Analysis

The layout of a source document may be complex at times. It can contain graphics, tables, and text arranged in different blocks. Before OCR program starts the recognition process, it needs to perform another step called “analysis.” This process distinguishes graphics from text, and identifies different text blocks based on the different fonts, sizes and printing directions. You can either do the analysis manually or have the software do it for you.

To manually analyze a selected page, first click the **Block Marking Tool** icon from the **Edit Toolbox** (Figure 4.1) to specify that you want to mark text blocks. Drag the mouse cursor to form a rectangle around the characters that belong together. They will be enclosed by a blue rectangle outline with black dots at the four corners. You can drag the black dots to resize the area, or drag the inside of the rectangle to move it as a whole. Create a new rectangle for each block of text that needs to be identified separately. Areas that are not enclosed will not go through recognition.

One interesting thing is that sometimes, even if you specify an area as one block, the software may get wise and decide to separate it into parts during the recognition process. This can happen when the program detects things like the spacing between two text rows being larger than normal. To prevent this, you can highlight the block and then select **Keep Block Intact** from the **Format** menu.

When a page contains graphics, you should mark these regions out as image blocks. This can be done in a similar way as marking a text block, but you should