Atiz Innovation, Inc. 2441 N. Beachwood Dr., #4 Los Angeles, CA 90068 USA www.atiz.com

How to Create an E-book

Using a flatbed scanner, an overhead scanner and a digital camera-based scanner.

Introduction

This application note gives an overview on methods to scan a book using three different kinds of book scanners. It compares and discusses the various advantages and disadvantages of using those methods.

Application

- Digital Library
- Imaging Service Bureau
- Bonded Document Digitization
- Book/Document Archives
- Content creation

Discussed methods

- Flatbed scanners
- Overhead Scanners
- V-shaped bookscanner

Procedure

Proper book digitization in a broad sense involves performing various tasks including administrative, logistical, and technical functions to name a few. As for the heart of the book digitization itself in a narrow sense, the procedure can be decomposed into three main processes. First, the content of the books specified for inclusion in the digitization project have to be turned into digital format via an image capturing device. Second, those captured images will normally have to go through software workflow for post image processing. The typical image operations include cropping, rotation, deskewing, resizing, format conversion, brightness and contrast adjustment as well as other image enhancing operations. Next, once the images are in acceptable quality, they will further go through text conversion using an OCR (optical character recognition) software package that will analyze the content of the images and turn them into text-editable, searchable files.

Scanning using a flatbed scanner

Average users typically rely on a flatbed scanner, a mature technology that has been available for about a decade and that was formerly a popular device for home and business use, to scan the content. Flatbed scanners were designed to scan flat material like sheets of paper so they are not always an optimal device for scanning other types of materials. Using this tool to scan a book can result in various problems that can be summarized as follows.

First, because books have a binding, and unless they are debinded for purpose of inserting into a a flatbed scanner or a sheetfed scanner, the body of the book will normally form curvature that causes dark and blurred areas around the gutter. Although it is possible to correct curvature and other related imperfections of the pages through software, this translates to extra processing time which can take long and the software correction is not completely reliable. Sometimes the software is not able to correct for some images and those would require humans to intervene and manually make correction so that those problematic pages come out having the same quality of other pages. Having said that, there are problems that cannot be corrected by software. For example, it is extremely hard to perform a

consistently reliable curvature correction on all the pages to make them naturally flat looking. Some hard-to-read content near the gutter represents the most difficult challenge for correction.

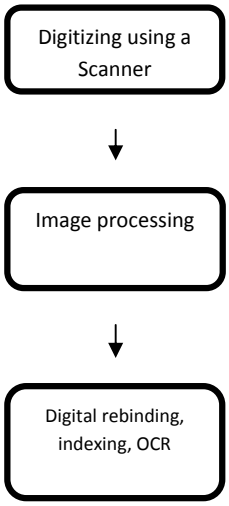
However, if it is ok to debind those books, the resulting debinded books that are in the form of sheets of paper will prove to be an easy scanning work for the operator. Using an automated sheet-fed scanner in such situation should result in the fastest scanning speed. The images will come out with flat-looking appearance that looks closest to the originals. However, it is not possible to debind all the bound books. Some of those books could be invaluable, hard-to-find books that are not allowed for any action that might impair the books.

The second step is processing the raw files (not to be confused with a RAW image file format) using image-processing software. The raw images may need to be cropped. The contrast and brightness may need to be adjusted. They may need to be rotated or deskewed, etc. Digitizers do this either page by page or with a batch process provided by some software.

Novice users often approach this stage by dealing with the images one by one at a time which is very time-consuming. Alternatively, all those image operations can be done in a batch process. The batch processes are quite useful. They set the parameters on the first page and apply them to all the pages. However, these features are not smart enough to detect the border correctly. The border areas can be the worst part of the image when using a flatbed scanner. Flatbed scanners provide a non-linear change of position. Therefore, you need to carefully place the book in the same position on the scanner each time or you will need to crop page by page after you are done.

The last process is binding the image into an Electronic book. The most popular file type used for e-books is the Adobe Acrobat (PDF) format. Users need to buy the professional version of Adobe Acrobat to create editable acrobat files. Users must also create a bookmark keyword using this software. Some users prefer to convert image into textusing Optical Character Recognition (OCR) software. A scanning resolution of at least 300 dpi is recommended for OCR.

E-Book process



Scanning method

Test images from the three types of book scanners are shown. These images are not processed by the image-processing software supplied by the manufacturer.

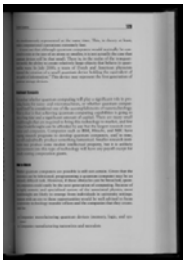


Figure 1
Image from a flatbed scanner

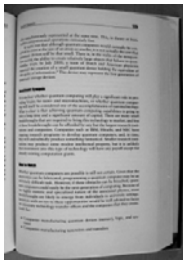


Figure 2
Image from an overhead scanner

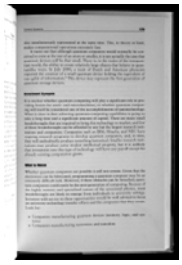


Figure 3
Image from a V-shaped scanner

Dark area problems

Flatbed scanners have problems with the dark area that occurs in the book gutter. Light from the flatbed scanner cannot reach the deepest part of the book. Even using a lens reduction type flatbed book scanner, which has a longer depth of field, cannot resolve this problem.. The image results are poor and the information in the middle cannot be recovered by the image-processing software.

The overhead scanner and v-shaped scanner had no problem with the dark area, but both require a larger area of workspace.

Book curvature problems

The image becomes distorted because the distance from the book to the sensors will reflect on the image.. When the object is not flat enough the image will be distorted or, when the distance is out of the DOF range, blurred..

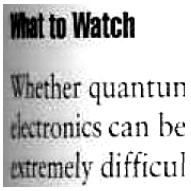


Figure 4
Image from flatbed scanner

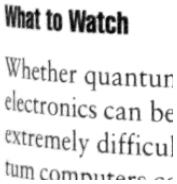


Figure 5
Image from overhead scanner

Figure 6
Image from v-shaped scanner

Using the overhead scanner the image is well illuminated. There is no dark area but this method has problems with book curvature. This can be corrected by re-processing using software, but it takes time.

The flatbed book scanner had only a few distorted areas. But that problem mixed with the dark-area problem made the image processing algorithm more complicated.

The V-shaped book scanner created the best image. It was well illuminated, and the book had no curvature distortion at the middle. However, it needs a large platform to scan.

Image processing

Cropping

After users scan the book into a digital file they need to crop the image to remove the unused space. Users can crop the image themselves if there are only a few images using the provided image-processing software from the manufacturer or other software such as Adobe Photoshop, ACD See, etc.,. However for the book scanning procedure, self-cropping is not recommended. The batch process is a better way. Users can set the parameters at the first page and wait for the software to process the rest. But keep in mind that even if the software is smart enough to detect the border of the book; users will need to scan images at the same position each time or check page by page that their software is detecting the right border and not misunderstanding border and image.

When using the flatbed scanner, users need to place the book in the same position everytime they scan and press the book down constantly.. Sometimes even doing this the images will move. This will make automatic cropping harder.

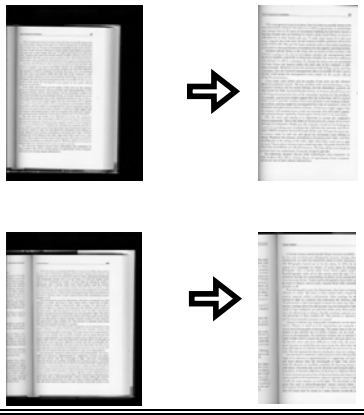


Figure 7

Both of the images use batch cropping.. The error pages occur when the position of the book was moved too much. Using a v-Shaped book scanner, the captured images are almost always at the same position.

Users can set the crop area at the first page and let the software do the other pages by itself.

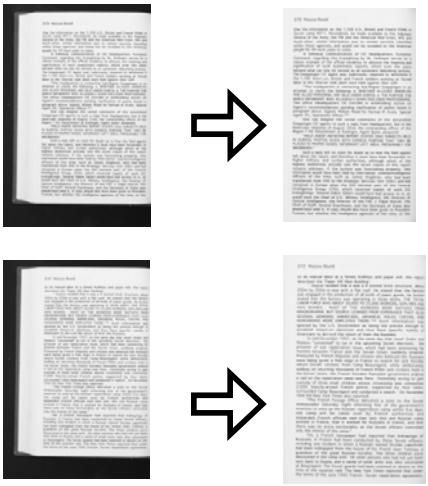


Figure 8

Even if users use the v-shaped book scanner the images move to the right a bit. This is caused by the thickness of the paper itself. We call this effect the "Margin crawl". Users need to change the crop setting every 100-200 pages depending on the thickness of the paper. This will only effect the horizontal size, the vertical size remains the same.

Figure 9

Procedure to prevent cropping problems

- Scan at the same position each time.
- Divide the images into three or more groups to crop (in the case of the v-shaped book scanner)

Bad alignment (Deskew)

If scanned images are not straight or unparallel with the border user need to deskew the image. If the scanned images are randomly tilted such as the image from the flatbed scanner, users will need to adjust the image deskew page by page or using the auto-deskew function. It will take a little bit of time to detect the tilt angle and deskew it back to being parallel with the border. This will increase the process time and decrease the image quality. The recommended procedure is scanning border parallel images before processing them..

	Flatbed scanner	***Overhead scanner***	***V-shaped scanner***
Image distortion	Moderate	High	Low
Dark area	High	Low	Low
Scan time	High	Low	Low
Size of machine	Low	Moderate	High

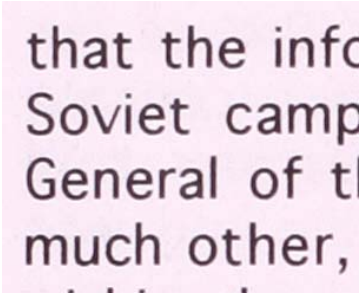
This problem occurs mostly with the flatbed scanners because it is very hard to straighten the book to the flatbed. If a user crops it without deskewing it, this will result in a bad cropping frame and decrease the e-book quality. Some OCR software is capable of understanding the tilted image but if you want a good OCR you better scan good images for them.

Procedure to prevent alignment problems

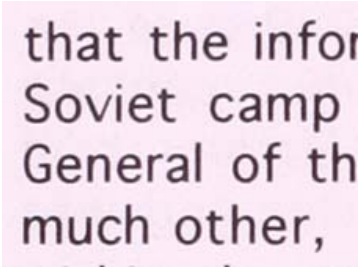
- Scan parallel images
- Use image processing software to deskew

Image size

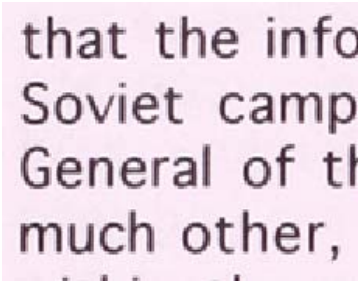
Most scanned image are in bitmap (BMP), JPEG (JPG) format. The bitmaps are raw image files that have a lot of detail because the file size is very large. The JPEG has smaller image format but even its smaller files are still large for digitizing a whole book. The quality of ook images decrease by changing the image size orthe color mode.



Original JPEG file (3.8 MB)



Resize to 70% (769KB)



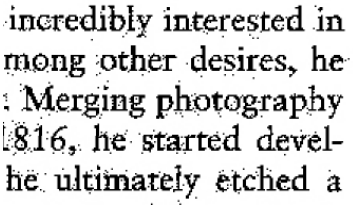
Resize to 50% (463KB)

In the example image shown you can see how dramatically resizing the file changes the image size. In the example, the original file from the scanner is a JPEG type with very high quality. Decreasing its size to 70% decrease the quality of the image to medium. The size can be reduced further to 20% of the original, but after you decrease to 50%the quality decreases but the file size does not decrease by much. Please make sure that images you resize are suitable for the OCR software.

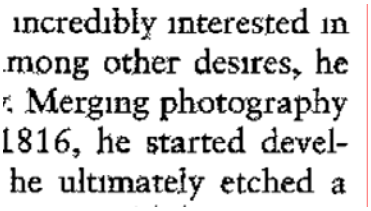
Noise (De-speckle)

For black and white format converting files from 24-bit color into black and white format will create some noise (speckles) on the image that cannot be removed completely. Even using the best high-quality scanner the image will still have noise on it. This noise can be remove by image processing software that has a de-speckle feature. The de-

speckle feature will remove the noise contained in the image without removing the text or print.


incredibly interested in
mong other desires, he
: Merging photography
816, he started devel-
he ultimately etched a

Raw image


incredibly interested in
mong other desires, he
: Merging photography
1816, he started devel-
he ultimately etched a

Processed image

The de-speckle feature will result in a clean image which contains only text. It will remove unwanted noise from the color conversion. It is best for OCR and book archive.

Black border

Images scanned from the scanner will contain a black border. This occurs especially with the flatbed scanner. The black border occurs on the large area because the images are illuminated poorly or are out of the cameras DOF. The black area can be removed by the software but the information contained in the shaded area is unrecoverable.

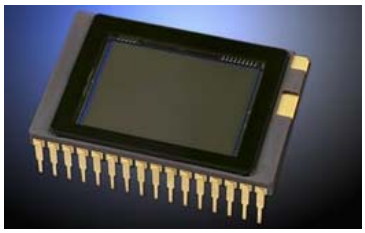
Resolution / Image sensor

Some digitizers are confused by the image sensors used in a digital camera and flatbed scanner. Both of the sensors use the same image sensor technology. The differences between them are that the flatbed scanner uses a linear image sensor and a digital camera uses a full-frame image sensor.

The linear image sensors are small and high resolution. They need to move along the document while it scans. This is a speed limitation for the flatbed. They cannot move fast or they would need to make a motorized part which would be more complicated and costly. The auto-feed scanner uses this kind of image sensor as well, but instead of driving the scanner they drive the document. The auto-feed scanner is faster than the flatbed scanner. This type of image sensor delivers a low-cost, small scanner solution.

The full-framed image sensor is bigger and has lower resolution when compared with the linear image sensor. It also requires a more complex lens to deliver a high-quality image.. Users who prefer high-quality imaging can use the professional digital SLR

camera. By using full-framed image sensors it needs more space and makes the digital camera scanning solution require a large space.



	Linear image sensor	*Full framed image sensor*
Size of sensor	Small	Large
Space required	Low	High
Time to scan	Long	Short
Cost	Low	High
Resolution	High	Low

Type of flatbed scanners

There are two types of flatbed scanners; the Lens Reduction type and the Contact Image Sensor type. Both types use linear images. It can be CCD or CMOS sensors depending on the manufacturer of the technology. Each type has benefits and drawbacks.

The CIS type provides a smaller and cheaper alternative because it has no lens and mirrors and needs no assembly. This type can be lowered to a one-inch thickness cause it uses a 1:1 optic coupler and is designed for digitizing flat paper where depth of field is not the problem.l

The Lens Reduction scanner is more expensive than the CIS. However, it delivers more resolution and has better image quality. It uses a lens assembly with a mirror to reflect inside the scanner. The depth of field of this type of scanner is larger than the CIS type which makes this scanner suitable for 3D objects including a book. Some scanner manufacturers claim they designed it specifically for the book.

Fatigue of the digitizing user

Imagine you are scanning two pounds of books (approximately 1000 pages.) on a flatbed scanner. You need to place each page on the same position and wait for the scanner head to slowly move along the bed. You are not allowed to move the book while it scans. You cannot reduce the resolution to speed up the scan rate because you must scan the book at 300 DPI. Yu also need to turn each of the 1000 pages one at a time.. This is inconvenient and time-consuming. That is why the flatbed scanner is not suitable for books.

Flatbed scanners use face-down scanning, this makes it easier for them to be light weight and have a small platform. These benefits however, are useless for book digitizing. Many early book scanners utilized face-up scanning, which made them easier to use and caused less fatigue to the user. They only had to turn the page by hand and wait for it to scan, no lifting was needed.

Resolution

The resolution is the amount of dots per unit of length. We usually use Dots per Inch (DPI) to measure resolution.. Resolution depends on both the image sensor size and the object size. Resolution of a flatbed scanner goes up to 2400 dpi (Optical resolution). This resolution is designed for film and small objects.

High resolution results in a larger file size, increased process time, increased storage size, etc,. Especially for the flatbed scanner, increasing the resolution will

increase the scan time. For OCR purposes you need only 300 DPI. You don't need to use the maximum resolution of the scanner.

For the digital camera we usually measure the resolution in the total amount of pixels (mega pixel). Cameras with more mega-pixels will result in better images. Users who use the digital camera as the capturing device need to convert their amount of pixels into DPI units.

Users need to check the effective image size with their digital camera manufacturer, and divide it with length or height of the object area.

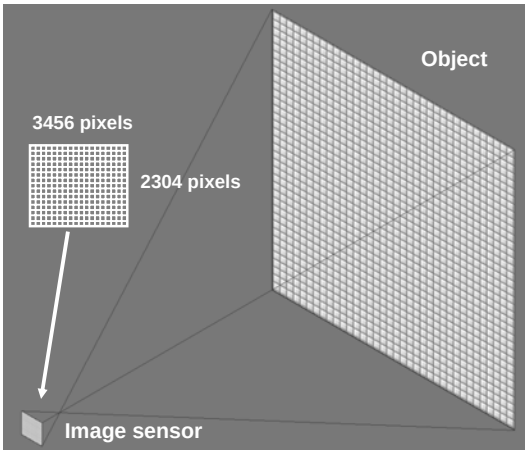


Figure 11

Example

Canon Rebel XT (EOS 350D) Digital SLR camera

Specification

- Total pixel 8.2 Megapixel
- Effective pixel = 3456 x 2304 pixel
- Aspect Ratio 4:3

Resolution:

- 2304 pixel / 8.27 inches = 279 pixel/inch. Approx 300 dpi.

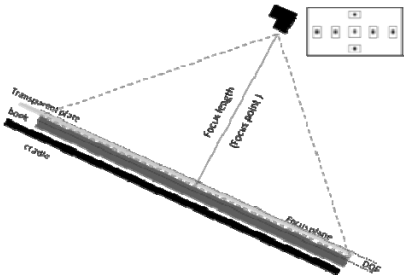
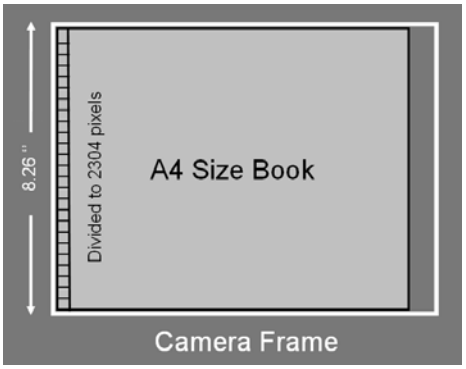


Figure 10

That means the sensor has 3456 usable pixels in the horizontal and 2304 usable pixels in vertical. The ration between horizontal and vertical is called the aspect ratio. This camera has a 3:2 aspect ratio. The DPI depends on the length between the lens and the object. A longer range equals a lower resolution.



Capturing A4 size book (8.26 x 11.7 inches)

For this example the aspect ratio between the book and the camera are not matched. The A4 size book has a lower ratio than the camera. This makes the book image fit the vertical but not fit on the vertical. Users need to calculate the resolution at the vertical of the book instead of calculating both sizes.

Speed

Many users ask us about the speed of the scanner. They always ask for a scanning rate (pages/hour). But there is another important rate. The time per page (seconds/page) in which the user needs to hold their book steady on the scanner. This increases the percentage of error that can occur when a book is accidentally moved while the scanner head is not finished with its task.

Activity	*Flatbed scanner*	*camera-based scanner*
Scanning/Capturing (300 dpi)	6	2
Turning page	2	0.5
Align the back book in line	2	0.5
Total time	10	3
Pages per hours	360	1200

By using the digital camera, scanning speed is not dependent on the resolution. The digital camera delivers faster than a scanner at same resolution, but it cannot increase the resolution at the same length.

Optical Character Recognition

We take the best images from each machine to OCR software. We are using Abbyy Finereader V.8 Professional Edition to test the images.

Result:

Using Flatbed scanner

Uncertain characters: 217

Total characters: 2267

Accuracy: 90.42 %

Using overhead scanner

Uncertain characters: 186

Total characters: 2214

Accuracy: 91.59 %

Using V-shaped book scanner

Uncertain characters: 8

Total characters: 2355

Accuracy: 99.66 %